AD_______________

GRANT NUMBER DAMD17-96-1-6312

TITLE: Using Neural Networks in Diagnosing Breast Cancer

PRINCIPAL INVESTIGATOR: David Fogel, Ph.D.

CONTRACTING ORGANIZATION: Natural Selection, Incorporated
La Jolla, CA 92037

REPORT DATE: September 1997

TYPE OF REPORT: Final

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 1997	**3. REPORT TYPE AND DATES COVERED** Final (1 Sep 96 - 31 Aug 97)

4. TITLE AND SUBTITLE

Using Neural Networks in Diagnosing Breast Cancer

5. FUNDING NUMBERS

DAMD17-96-1-6312

6. AUTHOR(S)

David Fogel, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Natural Selection , Incorporated
La Jolla, California 92037

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

Computational methods can be used to provide a second opinion in medical settings and may improve the sensitivity and specificity of diagnoses. In the current study, evolutionary programming is used to train neural networks and linear discriminant models to detect breast cancer in suspicious masses and microcalcifications using radiogarphic features and patient age. A cross validation protocol is used to train and atest the networks. ROC curves are used to assess the performance. Results indicate that a significant probability of detecting malignancies can be achieved at the risk of a small percentage of false positives. Typical areas under the ROC curves average 0.9 or better. The results compare well with others offered in the archive literature, while using an order-of magnitude fewer degrees of freedom in the neural classifiers. The research sets the stage for further investigation to automate the assessment of important indicators of breast cancer.

14. SUBJECT TERMS Breast Cancer, Mammography
artificial neural networks, evolutionary algorithms, computer-assisted diagnosis

15. NUMBER OF PAGES

75

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

✓ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

PI - Signature 18 SEP 1997
Date

Table of Contents

Introduction

Carcinoma of the breast is second only to lung cancer as a tumor-related cause of death in women. There are now more than 180,000 new cases and 45,000 deaths annually in the United States alone [1]. It begins as a focal curable disease, but it is usually not identifiable by palpation at this stage, and mammography remains the mainstay in effective screening. It has been estimated that the mortality from breast carcinoma could be decreased by as much as one-third if all women in the appropriate age groups were regularly screened.

Although there is currently considerable debate as to exactly at what age and how frequently to perform film-screen mammography, there is no doubt that large numbers of patients ideally do need to be evaluated by mammography at regular intervals. Conventional screening mammography generally includes, at a minimum, two views of each breast. The films have traditionally been evaluated to search for findings which would differentiate those considered normal from those showing abnormalities. Although the criteria for malignancy are reasonably well established, the application of such criteria is often quite subjective, and proper evaluation is a time consuming task for the radiologist, usually requiring a review of current and prior films (if available) by magnifying glass to search for the subtle microcalcifications that are so frequently an indicator of occult malignancy.

Computer technology offers many potential benefits to the radiologist. There is currently considerable intra- and inter-observer disagreement or inconsistencies in mammographic interpretation. This has led to an interest in the possibility of utilizing computerized pattern recognition algorithms, such as artificial neural networks, to assist in the decision-making required in the assessment of mammograms. Artificial neural networks have been demonstrated to be useful in many engineering pattern recognition applications and these techniques hold promise for improving the accuracy of determining those patients where further assessment and possible biopsy is indicated. Furthermore, there should also be an eventual cost savings when a reliable automated screening system can be developed. The successful development of a neural network that is capable of reliably assessing the potential for the existence of breast carcinoma based on radiographic features of mammograms would make the radiologist both more efficient and more effective.

Artificial Neural Networks for Pattern Recognition

Artificial neural networks (or simply *neural networks*) are computer algorithms loosely based on modeling the neuronal structure of natural organisms. They are stimulus-response transfer functions that accept some input and yield some output. They are typically used to learn an input-output mapping over a set of examples. For example, the input can be radiographic features from mammograms, with the output being a decision concerning the likelihood of malignancy.

Neural networks are parallel processing structures consisting of nonlinear processing elements interconnected by fixed or variable weights. They are quite versatile, for they can be constructed to generate arbitrarily complex decision regions for stimulus-response pairs. That is, in general, if given sufficient complexity, there exists a neural network that will map every input pattern to its appropriate output pattern, as long as the input-output mapping is not one-to-many (i.e., the same input having a variety of outputs). Neural networks are therefore well suited for use as detectors and classifiers. The classic pattern recognition algorithms require assumptions concerning the underlying statistics of the environment. Neural networks, in contrast, are nonparametric and can effectively address a broader class of problems [2].

Multi-layer perceptrons, also sometimes described as *feed forward networks*, are probably the most common architecture used in supervised learning applications (where exemplar patterns are available for training). Each computational node sums N weighted inputs, subtracts a threshold value and passes the result through a logistic (sigmoid) function. Single-layer perceptrons (i.e., feed forward networks consisting of a single input node) form decision regions separated by a hyperplane. If the input exemplars from the given different data classes are linearly separable, a hyperplane can be positioned between the classes by adjusting the weights and bias terms. If the inputs are not linearly separable, containing overlapping distributions, a least mean square (LMS) solution is typically generated to minimize the mean squared error between the calculated output of the network and the actual desired output. While single perceptrons can generate hyperplane boundaries, perceptrons with a hidden layer of processing nodes have been proven to be capable of approximating any measurable function [3], indicating their broad utility for addressing general pattern recognition problems.

Another versatile neural network architecture is the *radial basis function network.* Rather than partitioning the available data using hyperplanes, the radial basis function network clusters available data, often with the use of approximate Gaussian density functions. The network comprises an input layer of nodes corresponding to the input feature dimension, a single hidden layer of nodes with computational properties described below, and output nodes which perform linear combinations on the hidden nodes. Each connection between an input node and hidden node carries two variable parameters corresponding to a mean and standard deviation. Poggio and Girosi [4] proved that linear combinations of these near-Gaussian density functions can be constructed to approximate any measurable function. Therefore, like the multi-layer perceptron, radial basis functions are universal function approximators.

Given a network architecture (i.e., type of network, the number of nodes in each layer, the connections between the nodes, and so forth), and a training set of input patterns, the collection of variable weights determines the output of the network to each presented pattern. The error between the actual output of the network and the desired target output defines a response surface over an n-dimensional hyperspace, where there are n parameters (e.g., weights) to be adapted. Multi-layer feed forward perceptrons are the most commonly selected architecture and training these networks can be accomplished through a *back propagation* algorithm which implements a gradient search over the error response surface for the set of weights that minimizes the sum of the squared error between the actual and target values.

Although the use of back propagation is common in neural network applications, it is quite limiting. This procedure is mathematically tractable and provides guaranteed convergence, but only to a locally optimal solution. Even if the network's topology provides sufficient complexity to completely solve the given pattern recognition task, the back propagation method may be incapable of discovering an appropriate set of weights to accomplish the task. When this occurs, the operator has several options: (1) accept suboptimal performance, (2) restart the procedure and try again, (3) use ad hoc tricks, such as adding noise to the exemplars, (4) collect new data and retrain, or (5) add degrees of freedom to the network by increasing the number of nodes and/or connections.

Only this last approach, adding more degrees of freedom to the network, is guaranteed to give adequate performance on the training set, provided sufficient nodes and layers are available. Yet this also presents problems to the designer of the network, for any function can map any measurable domain to its corresponding range if given sufficient degrees of freedom. Unfortunately, such overfit functions generally provide very poor performance during validation on independently acquired data. Such anomalies are commonly encountered in regression analysis, statistical model building, and system identification. Assessing the proper trade-off between the goodness-of-fit to the data and the required degrees of freedom requires information criteria (e.g., Akaike's information criterion, minimum description length principle, predicted squared error, or others). By relying on the back propagation method, the designer almost inevitably accepts that the resulting network will not satisfy the maxim of parsimony, simply because of the nature of the training procedure itself. The problems of local convergence with the back propagation algorithm indicate the desirability of training with stochastic optimization methods such as simulated evolution which can provide convergence to globally optimal solutions.

Evolutionary Computation and Neural Networks

Natural evolution is a population-based optimization process. Simulating this process on a computer results in stochastic optimization algorithms that can often outperform classical methods of optimization when applied to difficult real-world problems. There are currently three main avenues of research in simulated evolution: evolutionary programming, evolution strategies, and genetic algorithms. The methods are broadly similar in that each maintains a population of trial solutions, imposes random changes to those solutions and incorporates the use of selection to determine which solutions to maintain into future generations and which to remove from the pool of trials. The methods differ in the types of random changes that are used and the methods for selecting successful trials. Fogel [5] provides a review of the similarities and differences between these procedures. The methods have been shown to possess asymptotic global convergence properties, and in some cases the techniques can be shown to have geometric rates of error convergence [5], making them attractive for function optimization problems.

The procedures generally proceed as follows. A problem to be solved is cast in the form of an objective function that describes the worth of alternative solutions. Without loss of generality, suppose that the task is to find the solution that minimizes the objective function. A collection (population) of trial solutions are selected at random from some feasible range across the available parameters. Each solution is scored with respect to the objective function. The solutions (parents) are then mutated and/or recombined with other solutions in order to create new trials (offspring). These offspring are also scored with respect to the objective function and a subset of the parents and offspring are selected to become parents of the next iteration (generation) based on their relative performance. Those with superior performance are given a greater chance of being selected than are those of inferior quality. Fogel [5] details examples of evolutionary algorithms applied to a wide range of problems, including designing neural networks, and the principal investigator (Dr. Fogel) has applied these techniques to real problems in pharmaceutical design, factory scheduling, and freeway onramp metering (e.g., [6,7]).

Designing neural networks through simulated evolution follows an iterative procedure:

1. A specific class of neural networks is selected. The number of input nodes corresponds to the amount of input data to be analyzed. The number of classes of concern (i.e., the number of output classification types of interest) determines the number of output nodes.

2. Exemplar data is selected for training.

3. A population of P complete networks is selected at random. A network incorporates the number of hidden layers, the number of nodes in each of these layers, the weighted connections between all nodes in a feed-forward or other design, and all of the bias terms associated with each node. Reasonable initial bounds must be selected for the size of the networks, based on the available computer architecture and memory.

4. Each of these "parent" networks is evaluated on the exemplar data. A payoff function is used to assess the worth of each network. A typical objective function is the mean-squared error between the target output and the actual output summed over all output nodes; this technique is often chosen because it simplifies calculations in the back propagation training algorithm. As evolutionary computation does not rely on similar calculations, any arbitrary payoff function can be incorporated into the process and can be made to reflect the operational worth of various correct and incorrect classifications. Information criteria such as Akaike's information criterion (AIC) [8] or the minimum description length principle [9] provide mathematical justification for assessing the worth of each solution, based on its classification error and the required degrees of freedom.

5. "Offspring" are created from these parent networks through random mutation. Simultaneous variation is applied to the number of layers and nodes, and to the values for the associated parameters (e.g., weights and biases of a multi-layer perceptron, weights, biases, means and standard deviations of a radial basis function network). A probability distribution function is used to determine the likelihood of selecting combinations of these variations. The probability distribution can be preselected *a priori* by the operator or can be made to evolve along with the network, providing for nearly completely autonomous evolution [5].

6. The offspring networks are scored in a similar manner as their parents.

7. A probabilistic round-robin competition is conducted to determine the relative worth of each proposed network. Pairs of networks are selected at random. The network with superior performance is assigned a "win." Competitions are run to a preselected limit. Those networks with the most wins are selected to become parents for the next generation. In this manner, solutions that are far superior to their competitors have a corresponding high probability of being selected. The converse is also true. This function helps prevent stagnation at local optima by providing a parallel biased random walk.

8. The process iterates by returning to step (5).

Application of Neural Networks for Diagnosis of Breast Cancer from Mammograms

While neural networks have been successfully applied to difficult engineering problems, their application to problems in medicine has been limited. More specifically, with respect to diagnosis of breast cancer, to date, only very recent research effort has been published in archived literature [10] [11] [12].

The investigation of [10] used 43 preselected features related to density, microcalcification, parenchymal distortion, skin thickening, correlation with clinical findings, and so forth. Data was taken from 133 textbook cases in [13]. For each mammogram, each of the selected features was rated by an experienced mammographer on a scale of 0-10, and this served as the vector input to a multi-layer perceptron neural network. The network possessed 10 hidden units in a single layer and a single output unit which was trained to yield a value of 0.0 for a benign case and 1.0 for a malignancy. Training was accomplished using back propagation. The results of this preliminary study and other described experiments indicated the suitability of this approach. By pruning the feature set to a more reasonable, smaller collection, the neural network was able to statistically significantly outperform attending radiologists and residents in assessing patterns of mammographic image features that are associated with benign and malignant lesions. There was no statistically significant difference between the performance of the network and the experienced mammographer used to rate each of the image features.

Floyd et al. [11] used back propagation on multi-layer perceptrons to predict breast cancer from mammographic findings from patients who were scheduled for biopsy. They used only eight input parameters (mass size, mass margin, asymmetric density, architectural distortion, calcification number, calcification morphology, calcification density, and calcification distribution) and each of these was parameterized more objectively (as opposed to subjectively) than it appears in [10]. There were 260 cases used for training and testing. There was no held-out training set; all of the exemplars were processed in a jackknife statistical procedure. After significant training, the results indicated that if a threshold value of 0.1 were used, 38 out of 168 benign cases and all 92 malignancies would be identified. The authors compared this performance to that of radiologists and suggested that these results were statistically significantly better than radiologists at a $P < 0.08$ level. Although their results do appear fairly impressive with regard to detecting malignancy, the number of false alarms appears rather high, and the statistical validity of the hypothesis test carried out can be questioned because the threshold of 0.1 was chosen after the authors reviewed the data and statistics were compiled on that same data. Thus the data did not reflect a random sample, but rather a biased sample. New data would have to be tested at the threshold value of 0.1 to provide a sound statistical assessment.

Wilding et al. [12] used back propagation on multi-layer perceptrons to assess both breast and ovarian cancer. Their procedure was similar to both [10] and [11] except that their input parameters consisted mainly of objective blood specimens and analyses (maximum of 10 total input parameters) from 104 patients. Unfortunately, Wilding et al. [12] reported that the neural network was able to "provide little improvement in the sensitivity of testing compared to the use of [the tumor marker] CA 15-3 only. Furthermore, it would appear that none of the networks appear to identify any worthwhile parameters or operating conditions with clinical utility."

As indicated from the above discussion, each of these investigations was limited in several important regards. First, multi-layer perceptrons may not generalize well on new data because they are partitioning networks rather than clustering networks. Second, the back propagation algorithm is well-known to lead to suboptimal convergence at locally optimal weights sets. Thus, a network trained with back propagation may require many more hidden nodes to train to a tolerable level of error than are actually required. This excess in degrees of freedom subsequently hinders the generalization properties of the network, as it essentially overfits noise in the data. These concerns were specifically discussed in [11] and [12]. The most effective methods employed in these investigations for limiting the number of nodes and network parameters were based on sensitivity analysis and ad hoc pruning. Sensitivity analysis is problematic on nonlinear transfer functions (such as neural networks) and ad hoc pruning can be largely unproductive. Despite directly mentioning concerns about overfitting their data, Floyd et al. [11] found that their best performance occurred when using 177 weights (16 hidden nodes), but they only had 260 samples. Wilding et al. [12] used networks with as few as 38 weights and as many as 132 weights, and despite having 104 samples were still unable to generate satisfactory performance. Even if the blood statistics that were being used were not particularly relevant to the classification task at hand, the failure to find suitable networks with more parameters than data indicates the limitations of the training method and selected neural architecture.

The current investigation has improved upon the performance generated in the efforts of [10], [11], [12], and other more recent efforts documented in the archive literature (see Body section). The current year-long study has used both radial basis functions (receptive fields) and multi-layer perceptrons were used in the classification of mammograms. Evolutionary computation was used to train the architectures to classify data collected by Dr. Eugene Wasson (domain expert) in accordance with approved protocols. The results have demonstrated a statistically significant improvement over the state-of-the-art while using an order-of-magnitude simpler neural classifier and only 13 input features. These results indicate potential for advancing the research toward the practical implementation of image processing filters to serve as inputs to a neural classifier, and eventually toward the application of a commercially available product that incorporates computer-assisted detection and classification of suspicious features in film screen or digital mammograms.

Body

Data were acquired in essentially four phases under the purview of Dr. Eugene C. Wasson, III, M.D. of Maui Memorial Hospital in Wailuku, Maui, Hawaii, who has over 20 years of experience in radiology and breast cancer diagnosis. The surrounding area of the study provides a small community with a population of approximately 120,000 people and a single acute care hospital. In 1993, Maui Memorial Hospital had between 50 and 60 positive breast biopsies for carcinoma and had at least five times as many negative biopsies, all with mammography. The size and composition of the community offers a measure of control that could not be expected at a hospital that serves a major metropolitan city, although expanding the investigation to a larger data base from just such hospitals now appears appropriate. The demographics for the data base reflect the local community of 24% part Hawaiian, 22% Caucasian, 17% Japanese, 17% mixed non-Hawaiian of miscellaneous races, 16% Filipino, and other cases being at or considerably below 2% representation. Computer programs were developed and executed on the Natural Selection, Inc. desktop machines and the SP-2 at the Maui High Performance Computing Center (facilitated by Edward M. Boughton).

Phase I

Data were collected by assessing film screen mammograms in light of a set of radiographic features as determined by the domain expert (Wasson). The features selected paralleled those of [8] with some important modifications. Under the system of [11], certain features were described as lying on a continuum when it appeared more useful to rate these features independently. For example, Floyd et al. [11] rated mass margin with six categories: (1) no mass (value 0.0), (2) well circumscribed (value 0.2), (3) microlobulated (value 0.4), (4) obscured (value 0.6), (5) indistinct (value 0.8), and (6) and spiculated (1.0). In contrast, the current parameterization rated the five categories of masses (see (2)-(6) in Table 1) in four levels $\{0,1,2,3\}$ as none, low, medium, and high. The complete set of radiographic features used appears in Table 1. In addition, the age of the patient was considered, leading to a total of 13 input features. In the first phase, 96 cases were analyzed. In all cases (across all phases), the indication of malignant or benign condition was confirmed by open surgical biopsy of the area of concern with the associated pathology indicating whether or not a malignant condition had been found. Of the 96 cases, 62 were associated with biopsy-proven malignancy and 34 cases were indicated to be negative by biopsy.

These data were processed using two forms of neural networks: (1) multilayer perceptrons and (2) receptive fields (radial basis functions). Each network architecture was restricted to two hidden nodes, with a linear output node, resulting in 31 adjustable weights (see Figure 1). Two hidden nodes were chosen as being the most parsimonious choice that still offered a nonlinear discriminant function. The perceptron network used a sigmoid filter on each hidden node of $f(\beta) = (1 + \exp(-\beta))^{-1}$, where β is the sum of a bias term and the dot product of the input feature vector and the associated weight vector. The receptive field network used a Gaussian filter on each hidden node of $f(\beta) = (2\pi)^{-0.5}\exp(-\beta^2)$. Evolutionary programming was used to train the networks in a leave-one-out cross validation procedure.

Specifically, for each complete cross validation where each sample pattern in turn was held out for testing then replaced in a series of 96 separate training procedures, a population of 250 networks of the chosen architecture were selected at random by sampling weight values from a uniform random variable distributed over [-0.5,0.5]. Each weight set (i.e. candidate solution) also had an associated self-adaptive mutational vector used to determine the random variation imposed during the generation of offspring networks (described below). Each of the self-adaptive parameters was initialized to a value of 0.01. Each weight set was evaluated based on how well the network classified the 95 remaining available training patterns (with one "left out" for testing), where a malignant condition was assigned a target value of 1.0 and a benign conditions was assigned a target of 0.0. The performance of each network was determined as the sum of the squared error between the output and the target value taken over the 95 available patterns.

After evaluating all existing (parent) networks, the 250 weight sets were used to generate 250 offspring weight sets (one offspring per parent). This was accomplished in a two-step procedure. For each parent, the self-adaptive parameters were updated as:

$$\sigma'_i = \sigma_i \exp\left(\tau N(0,1) + \tau' N_i(0,1)\right) \tag{1}$$

where $t = \frac{1}{\sqrt{2n}}$, $t' = \frac{1}{\sqrt{2\sqrt{n}}}$, $N(0,1)$ is a standard normal random variable sampled once for all $n = 33$ parameters of the vector σ, and $N_i(0,1)$ is a standard normal random variable sampled anew for each parameter. The settings for τ and τ' have been recommended as robust in [14]. These updated self-adaptive parameters were then used to generate new weight values for the offspring according to the rule:

$$x'_i = x_i + \sigma'_i C \tag{2}$$

where C is a standard Cauchy random variable

$$f(y) = \frac{1}{\pi\left(1 + y^2\right)}, -\infty < y < \infty \tag{3}$$

(determined as the ratio of two independent standard Gaussian random variables). Traditional methods in evolutionary programming and evolution strategies have relied on Gaussian mutation, however recent research in [15], [16], and others, have suggested a possible benefit to using Cauchy variation because it has a greater probability to generate longer jumps than the Gaussian. This offers a greater chance of escaping local optima on a error surface at the expense of poorer fine tuning. Initial observations with both Gaussian and Cauchy mutations on the existing data appeared to favor the Cauchy distribution, however a more careful analysis remains for future study. All of the offspring weight sets were evaluated in the same manner as their parents.

Selection was applied to eliminate half of the total parent and offspring weight sets based on their observed error performance. Following typical methods in evolutionary programming [5], a pairwise tournament was conducted where each candidate weight set was compared against a random sample from the population. The sample size was chosen as 10 (following [5], a greater sample size indicates more stringent selection pressure). For each of the 10 comparisons, if the weight set had an associated classification error score that was lower than the randomly sampled opponent it received a "win." After all weight sets had participated in this tournament, those that received the greatest number of wins were retained as parents of the next generation.

This process was iterated for 100 generations, this being believed to be sufficient, whereupon the best available network as measured by the training performance was used to classify the held out input feature vector. The result of this classification was recorded (i.e., the output value of the network and the associated target value) and the process was restarted by replacing the held out vector and removing the next vector in succession until all 96 patterns had been classified. Each complete series of cross validation was repeated 10 times for both the multilayer perceptron and receptive field networks.

A typical rate of optimization in each training run is shown in Figure 2. The overall training error often fell as a nearly linear function of the number of generations without saturation. This suggests that further training time might have been warranted. The probability of detection, P(D), and false alarm (false positive), P(FA), vary as a function of the discrimination threshold applied to the output of the networks. As the threshold value is lower, the network can correctly identify a greater number of cancers, but this comes at the expense of a higher false alarm rate. Conversely, the false alarm rate can be lowered by raising the threshold value, but this in turn decreases the sensitivity of the procedure.

The effectiveness of the classification procedures can be assessed using receiver operating characteristic (ROC) analysis, where the probability of detecting a malignancy is traded off as a function of the likelihood of a false positive result. Typical ROC curves for the multilayer perceptron and receptive field networks are shown in Figure 3. The area under the curve provides a useful measure for comparison.

To compare the effectiveness of the multilayer perceptron architecture with the receptive field, the area under the ROC curve for each of the 10 trials of cross validation with each method was estimated. This was accomplished by performing a polynomial regression of at least third order to the available samples of fraction of false alarms versus fraction of detections in each ROC curve. Regression models were determined by choosing the lowest order polynomial that provided (1) an R^2 value of at least 0.99, and (2) was non-decreasing over false alarm rates from zero to one (see Figure 4). The models were integrated over the range [0,1] to computer the desired areas. The mean area under the ROC curve (and standard deviation) for the perceptron and receptive field networks, respectively, were 0.787611±0.022346 and 0.739060±0.035574. Under a two-sample mean *t*-test (which assumes populations of normally distributed values), these data indicate statistically significant evidence in favor of the perceptron (sigmoid) networks ($P < 0.01$).

Phase II

Further data were collected and attention was divided between suspicious features of mass and microcalcifications. In all, 112 cases of suspicious breast mass (63 biopsy-proven malignant, 49 negative by biopsy) were examined. A similar method was applied to classify these data using cross validation. These data were processed using a simple feedforward ANN restricted to two hidden sigmoid nodes (following the maxim of parsimony, this being the simplest architecture that can take advantage of the nonlinear properties of the nodes), with a single linear output node, resulting in 31 adjustable weights.

The evolution was iterated for 200 generations (as compared with 100 in prior effort — the rate of error reduction indicated that further training might be useful), whereupon the best available network as measured by the training performance was used to classify the held-out input feature vector. The result of this classification was recorded (i.e., the output value of the network and the associated target value) and the process was restarted by replacing the held-out vector and removing the next vector in succession until all 111 patterns had been classified.

Each complete cross validation was repeated 16 times, with different randomly selected populations of initial weights, to determine the reliability of the overall procedure. A typical rate of optimization in each training run is shown in Figure 5. The probability of detection, P(D), and false positive, P(FP), vary with the discrimination threshold applied to the output of the networks. As the threshold value is lowered, the network can correctly identify a greater number of cancers at the expense of a higher false positive rate. Conversely, the false positive rate can be lowered by raising the threshold value, but this in turn decreases the sensitivity of the procedure.

The effectiveness of the classification procedures can be assessed using receiver operating characteristic (ROC) analysis, where the probability of detecting a malignancy is traded off as a function of the likelihood of a false positive. A typical ROC curve for the 16 trials is offered in Figure 6. The mean area $\bar{A}_Z$ (determined using polynomial splines) was 0.8982 with a standard error of $s_{\bar{A}_Z} = 0.0098$. The best network achieved $A_Z = 0.9345$.

The average performance of the evolved ANNs in terms of A_Z is comparable to that of [10] and [17] which also used mammographic features interpreted by a radiologist. The ANN in [17], which used 18 input features (both radiographic and clinical) and possessed 10 hidden nodes, yielded a specificity of 0.62 at a sensitivity of 0.95. By comparison, radiologists attained only a 0.3 specificity on the same data. The evolved networks in the current study yielded a mean specificity of 0.6187±0.0285 at 0.95 sensitivity. Although this result is almost identical to the performance offered in [14], the evolved networks were more parsimonious models (about an order of magnitude fewer degrees of freedom), and may therefore offer greater generalizability while requiring less computational effort.

Phase III

With additional data collected, comparisons were made between neural (i.e., nonlinear) and linear classifiers. The 139 cases of suspicious breast mass (79 cases were associated with a biopsy-proven malignancy, while 60 cases were indicated to be negative by biopsy) were processed using two input-output models: (1) a simple feedforward ANN restricted to two hidden sigmoid nodes (following the maxim of parsimony, this being the most simple architecture that can take advantage of the nonlinear properties of the nodes), with a single linear output node, resulting in 31 adjustable weights, and (2) a linear classifier (created by reducing the number of hidden nodes in the above ANN to one). Evolutionary programming was used to train both the ANNs and the linear discriminant classifier in a leave-one-out cross validation procedure (the optimum parameters for the linear models could have been determined by direct calculation, however, using the existing evolutionary optimization software facilitated the cross validation of these models).

Each complete cross validation was repeated 16 times for the ANNs and 10 times for the linear classifier, with different randomly selected populations of initial weights, to determine the reliability of the overall procedure. Typical rates of optimization in each training run are shown in Figure 7. The probability of detection, P(D), and false positive, P(FP), vary with the discrimination threshold applied to the output of the models. As the threshold value is lowered, the models can correctly identify a greater number of cancers, but this comes at the expense of a higher false positive rate. Conversely, the false positive rate can be lowered by raising the threshold value, but this in turn decreases the sensitivity of the procedure.

The effectiveness of the classification procedures can be assessed using receiver operating characteristic (ROC) analysis, where the probability of detecting a malignancy is traded off as a function of the likelihood of a false positive. Typical ROC curves for the 16 trials involving ANNs and 10 trials with linear classifiers is offered in Figure 8. The area under an ROC curve, often denoted A_Z, provides a useful measure for assessing performance. The mean area $\bar{A}_Z$ (determined using polynomial splines of maximum 9th order) for the ANNs was 0.9290 with a standard error of $s_{\bar{A}_Z} = 0.0052$. The best network achieved $A_Z = 0.9657$. The mean area $\bar{A}_Z$ for the linear classifiers was 0.9187 with a standard error of $s_{\bar{A}_Z} = 0.0048$. The best linear classifier achieved $A_Z = 0.9397$. A *t*-test indicated no statistically significant difference ($P > 0.1$) in performance between the ANNs and linear classifiers (note that the assumptions for the test may not hold given the sample sizes).

The average performance of the evolved ANNs and linear classifiers in terms of A_Z is slightly better than observed in [18] and comparable to that of [10, 17], which also used mammographic features interpreted by a radiologist. The ANN in [17], which used 18 input features (both radiographic and clinical) and possessed 10 hidden nodes, yielded a specificity of 0.62 at a sensitivity of 0.95. By comparison, radiologists attained only a 0.3 specificity on the same data. The evolved ANNs and linear classifiers in the current study yielded mean specificities of 0.7289±0.0346 and 0.6610±0.0413, respectively. Again, these values do not indicate a statistically significant difference ($P > 0.1$), however the performance of the evolved ANNs is statistically significantly better than that offered in [17] ($P < 0.05$).

Phase IV

The available data base of cases was brought to over 200 with over 100 individual examples of malignant and benign conditions at the beginning of August, 1997. These data have been processed in a manner similar to that described above. A typical ROC curve from 16 trials with 158 suspicious breast masses appears in Figure 9. The mean area $\bar{A}_Z$ (determined using polynomial splines of maximum 9th order) for the ANNs was 0.9196 with a standard error of $s_{\bar{A}_Z} = 0.0040$. The best network achieved $A_Z = 0.9487$. Figure 10 shows the cell point chart of the network output for benign and malignant cases for the 16th trial. The graph indicates good separation between the two classes. Figure 11 shows the histograms of network outputs for both classes, and again there appears to be good separation.

Individual trials were also conducted using the complete set of suspicious breast masses and microcalcifications. Figure 12 indicates that the area under the ROC is not as great as when attention in limited only to suspicious masses. This suggests an opportunity to improve the features used for assessing microcalcifications. Comparisons were conducted by varying the number of hidden nodes from 1 (a linear system) through 5, to determine if a nonlinear classifier (i.e., the ANN) provided any benefit to classification. Figure 13 shows that indeed, the use of two hidden nodes increases the overall classification performance in generalization; however, additional nodes to not lead to improved performance. This provides some justification for the use of small neural networks to perform the required classification.

Conclusions

Statement of Contract Performance

Effort on this one-year grant has completed the statement of work as described in the contract proposal. In particular, Natural Selection, Inc. has facilitated the collection of an appropriate data base of assessed mammographic and related features with corresponding identification of the presence or absence of malignancy (using biopsy proven cases). Statistical and graphical analysis of these features was performed to identify any obvious relationships between the features and diagnosis, however none was found. Training and test sets were created, however, following in line with prior literature, the use of separate training and test sets was deferred in favor of the use of cross validation techniques.

Evolutionary programs for training multilayer perceptrons and radial basis function (receptive field) networks were written. Using a squared error criterion, ultimately, over 200 cases were processed and performance was assessed. The results indicate a significant advance over prior attempts to generate neural classifiers of mammographic data. Natural Selection, Inc. has also identified a potential organization (Qualia Computing, Inc.) that could assist in further testing and transition of developed software, and this has led to submission of two proposals for further research and development. Efforts continue with regard to identifying the suitability of digital or scanned mammograms (Qualia may be able to supply these for future efforts). Finally, all relevant efforts have been documented in this final report.

Significance of Results and Implications

One criticism of the use of ANNs in medical diagnoses is that they are black box methods, and in general are not "explainable" [19]. The success of small ANNs in diagnosing breast cancer, as observed here, offers the promise that suitable explanations for the network's behavior can be induced, perhaps leading to a greater acceptance by physicians and ultimately a useful tool.

It would appear that training by simulated evolution or other stochastic methods is key to developing these parsimonious networks. Under the more common gradient-based training method of error back propagation, the search for appropriate ANN weight sets can stagnate at local optima. These can be overcome by adding additional nodes and weights, but the resulting networks are no longer as parsimonious as may be possible. Evolutionary algorithms offer the potential for overcoming multiple optima on the error response surface, as well as simultaneously adjusting the ANN topology. Further, the evolutionary training method can be used regardless of the payoffs for correct and incorrect classifications, which may be important in trading off the costs of sensitivity and specificity.

The current research demonstrates that there is sufficient information in the features selected to provide reasonable detection and diagnosis of malignant and benign conditions based on radiographic information and patient age. The following steps would be useful in furthering this effort toward a practical impact on women's health care:

1. The current sample size of mammograms should be increased continually.
2. Additional radiographic features should be considered, particularly with regard to assessing microcalcifications.
3. A variety of experts in mammography should take an active role in assessing the chosen features and providing a baseline for comparison to the performance of evolving neural classifiers.

4. Image processing algorithms should be developed to generate feature ratings that are similar to those offered by qualified experts in mammography when facing the same film screens.
5. Methods for incorporating additional relevant information from previous patient screening, patient history, or related individual factors should be developed and tested.
6. Collaborations with other groups performing similar efforts to automate diagnosis of breast cancer from film screen and/or digital mammograms should be encouraged as there is a significant potential for evolutionary optimization to provide added value to these efforts.

Surely these are but some of the possible avenues to pursue. The current effort has indicated that simple neural classifiers can offer the radiologist a potentially useful tool. Natural Selection, Inc. looks forward to opportunities to further the utility of this approach.

References

1. C.C. Boring, T.S. Squires, and T. Tong (1993) "Cancer statistics," *CA: Cancer Journal for Clinicians*, vol. 43, pp. 7-26.

2. R.P. Lippmann (1987) "An Introduction to computing with neural nets," *IEEE ASSP Magazine*, April, pp. 4-22.

3. K. Hornik, M. Stinchcombe and H. White (1989) "Multilayer feedforward networks are universal approximators," *Neural Networks,* Vol. 2, pp. 359-366.

4. T. Poggio and F. Girosi (1990) "Networks for approximation and learning," *Proc. of the IEEE,* Vol. 78:9, pp. 1481-1497.

5. D.B. Fogel (1995) *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, IEEE Press, Piscataway, NJ.

6. D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, and S.T. Freer (1995) "Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming," *Chemistry and Biology*, Vol 2:5, pp. 317-324.

7. J. McDonnell, D.B. Fogel, L.J. Fogel, C. Rindt, and W. Recker (1995) "Evolving optimal ramp control rules," *International Journal of Expert Systems*, Vol. 8:3, pp. 287-308.

8. D.B. Fogel (1991) "An information criterion for optimal neural network selection," *IEEE Trans. Neural Networks*, Vol. 2:5, pp. 490-497.

9. D.B. Fogel and P.K. Simpson (1993) "Experiments with evolving fuzzy clusters," *Proc. of 2nd Ann. Conf. on Evolutionary Programming*, D.B. Fogel and W. Atmar (eds.), Evolutionary Programming Society, La Jolla, CA, pp. 90-97.

10. Y.Z. Wu , M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schmidt, and C.E. Metz (1993) "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, Vol. 187:1, pp. 81-87.

11. C.E. Floyd, J.Y. Lo, A.J. Yun, D.C. Sullivan, and P.J. Kornguth (1994) "Prediction of breast cancer malignancy using an artificial neural network," *Cancer*, Vol. 74, pp. 2944-2998.

12. P. Wilding, M.A. Morgan, A.E. Grygotis, M.A. Shoffner, and E.F. Rosato (1994) "Application of backpropagation neural networks to diagnosis of breast and ovarian cancer," *Cancer Letters*, Vol. 77, pp. 145-153.

13. L. Tabar and P.B. Dean (1985) *Teaching Atlas of Mammography*, 2nd ed., Thieme-Stratton, New York.

14. T. Bäck and H.-P. Schwefel (1993) "An overview of evolutionary algorithms in parameter optimization," *Evolutionary Computation*, Vol. 1:1, pp. 1-24.

15. X. Yao and Y. Liu (1996) "Evolving artificial neural networks through evolutionary programming," *Evolutionary Programming V: Proceedings of the Fifth Annual Conference on Evolutionary Programming*, L.J. Fogel, P.J. Angeline, and T. Bäck (eds.), MIT Press, Cambridge, MA, pp. 257-266.

16. N. Saravanan and D.B. Fogel (1997) "Multi-operator evolutionary programming," *Evolutionary Programming VI: Proceedings of the Sixth Annual Conference on Evolutionary Programming*, P.J. Angeline, R.C. Eberhart, R.G. Reynolds, and J.R. McDonnell (eds.), Springer, Berlin, pp. 215-221.

17. Baker, J.A., Kornguth, P.J., Lo, J.Y., Williford, M.E., and Floyd, C.E. (1995) Breast cancer: Prediction with artificial neural networks based on BI-RADS standardized lexicon. *Radiology*, 196, 817-822.

18. D.B. Fogel, E.C. Wasson, E.M. Boughton, and V.W. Porto, "A step toward computer-assisted mammography using evolutionary programming and neural networks," *Cancer Letters*, in press, 1997.

19. Kahn, C.E. (1996) "Decision aids in radiology," *Imaging and Information Management: Computer Systems for a Changing Health Care Environment,* 34:3, 607-628.

Table 1. The features and rating system used for assessing mammograms in the current study. Assessment was made by the domain expert (Wasson).

1. Mass size: either zero or in mm.
2. Mass margin: (each subparameter rated as none (0), low (1), medium (2), or high (3))
    (a) Well circumscribed
    (b) Microlobulated
    (c) Obscured
    (d) Indistinct
    (e) Spiculated
3. Architectural distortion: none or distortion
4. Calcification number: none (0), < 5 (1), 5-10 (2), or > 10 (3).
5. Calcification morphology: none (0), not suspicious (1), moderately suspicious (2), or highly suspicious (3)
6. Calcification density: none (0), dense (1), mixed (2), faint (3)
7. Calcification distribution: none (0), scattered (1), intermediate (2), clustered (3)
8. Asymmetric density: either zero or in mm.

Figure 1. The design used for processing data, both for the multilayer perceptron and receptive field neural architectures. Input data are weighted in connections to the two hidden nodes. Each hidden node passes the sum of a bias term (not shown) and the dot product of the weights and inputs through a nonlinear filter. The filter is $f(ß) = 1/(1+e^{-ß})$ for the multilayer perceptron, and $f(ß) = (2\pi)^{-0.5}e^{-ß^2}$ for the receptive field. The output node is a linear filter which performs the sum of a bias term with the dot product of filtered hidden nodes and their associated weights. There are 31 weights given 13 inputs.

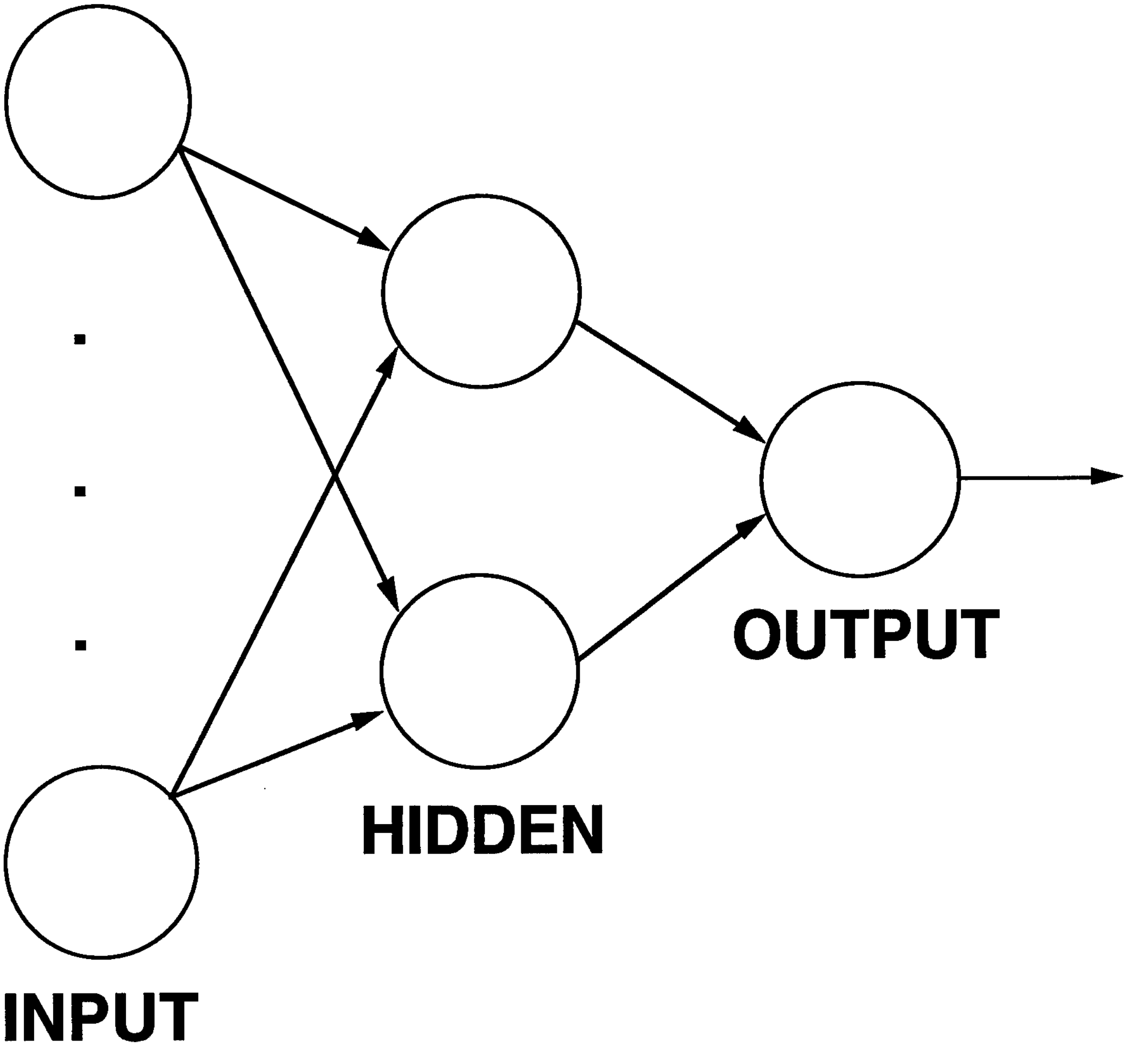


Figure 2. A typical rate of optimization in an evolutionary training of the neural networks on 95 patterns (one held out) over 100 generations. The error of the best member (weight set) in the population is seen to decrease nearly linearly as function of the number of generations. With further training, the observed best error would eventually saturate at an asymptote. The error is taken as the sum of the squared difference between the target value and the realized output from the network generated as a result of each input pattern.

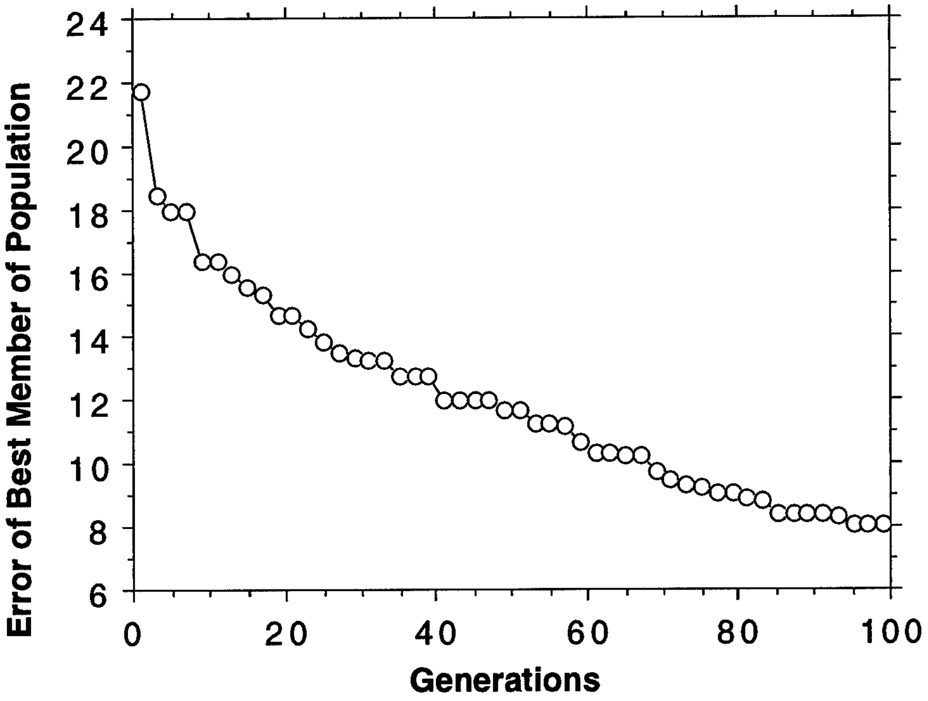
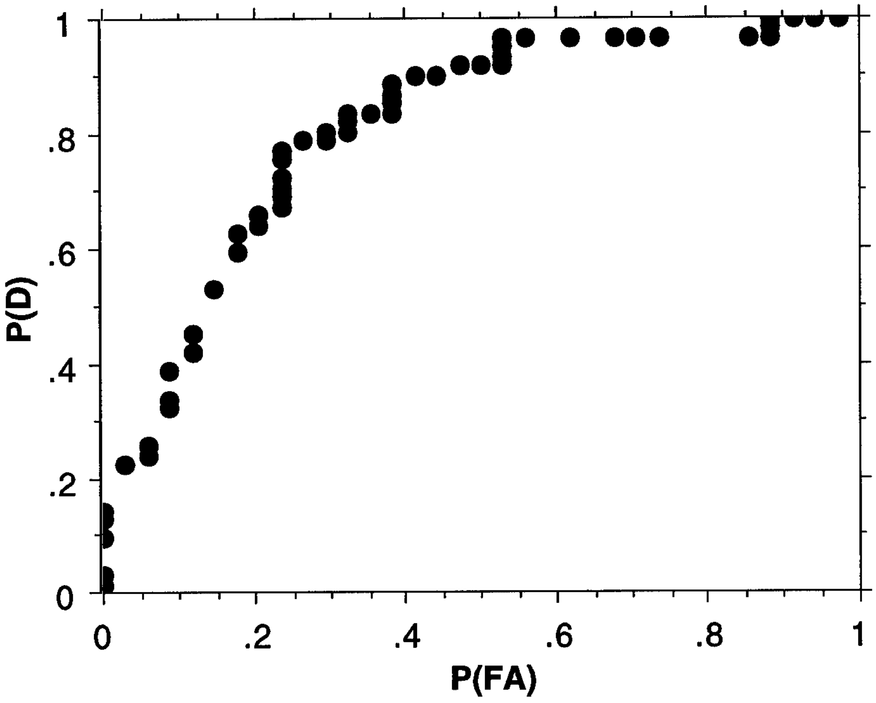
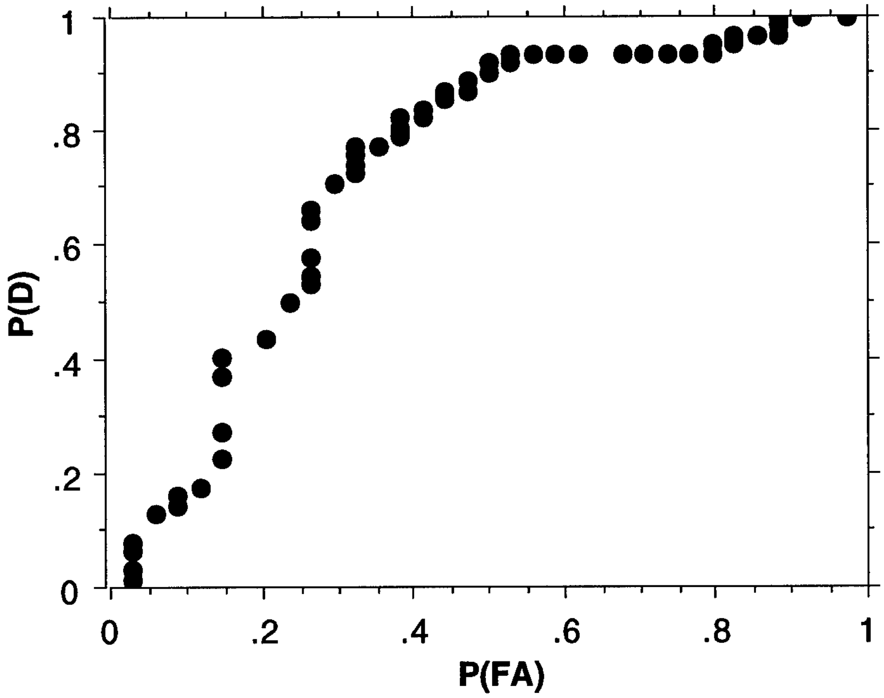


Figure 3. Typical ROC curves for the (a) perceptron and (b) receptive field neural networks. As the probability of a false alarm (i.e., indication of malignancy when none is present) increases, so does the probability of detection.



(a)



(b)

Figure 4. An example of using polynomial regression to estimate the ROC curve based on the observed pairs of probabilities for false alarm and detection. The equation shown is $P(D) = 4.742 \times P(FA) - 8.897 \times P(FA)^2 + 7.452 \times P(FA)^3 - 2.295 \times P(FA)^4$. The goodness-of-fit is $R^2 = 0.999$. Note that the regression equation is constrained to pass through the origin. Regressions were conducted for each of the 10 complete cross validation studies with both the perceptron and receptive field networks in order to determine the area under the approximate ROC curve.

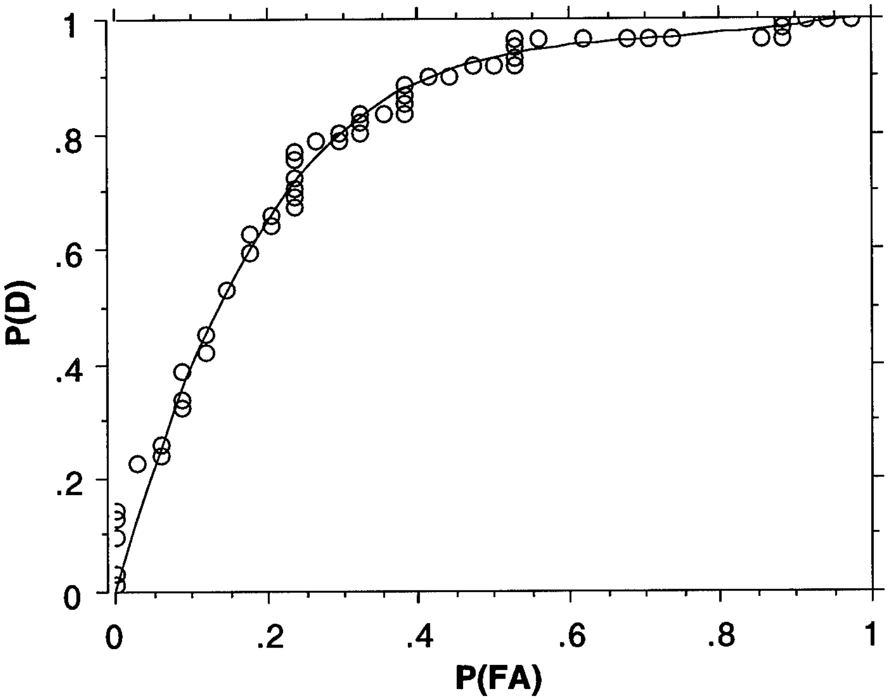


Figure 5. Typical optimization performance using simulated evolution to train the ANN. The graph depicts the sum of squared error (SSE) of the best network in the population as a function of the number of generations. Training was performed over 111 patterns, with one pattern held out for testing in cross validation. The sufficiency of the number of generations is indicated as the learning curve approaches an asymptote.

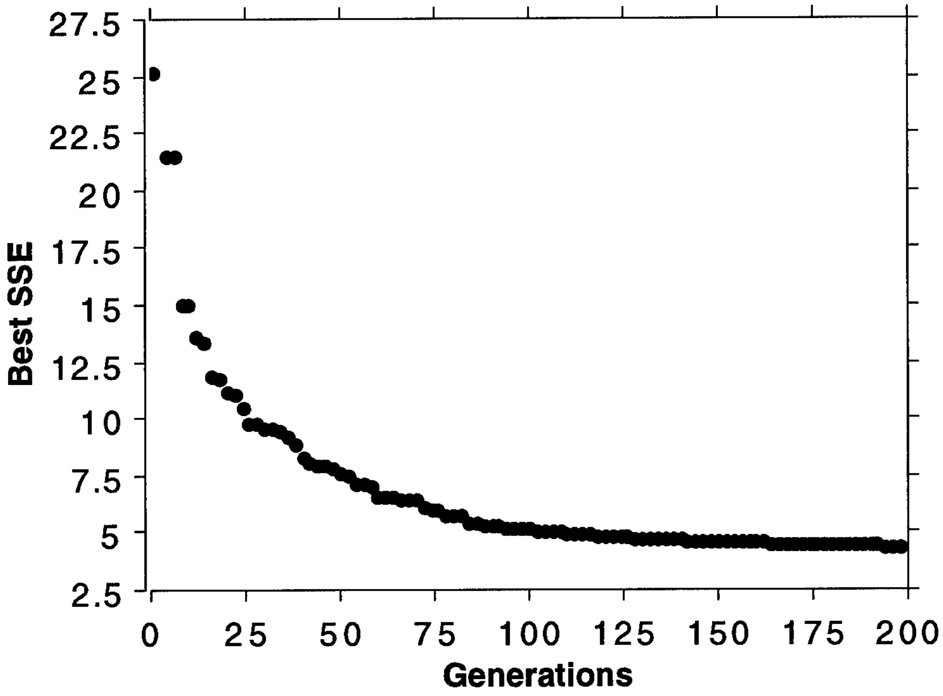


Figure 6. A typical ROC curve (raw data) generated in one complete cross validation where each of 112 patterns was classified in turn, based on training over the remaining 111 patterns. Each point represents the probability of detection, P(D), and probability of false positive, P(FP), that is attained as the threshold for classifying a result as malignant is increased systematically over [0,1].

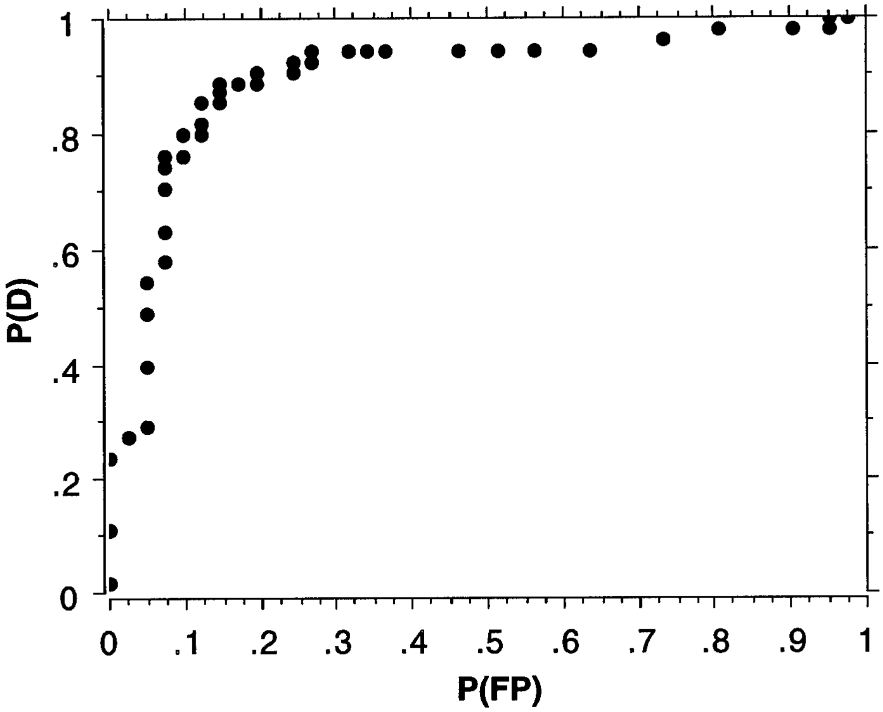


Figure 7. Typical optimization performance using simulated evolution to train (a) the ANN, (b) the linear classifier (an ANN with only one hidden node). The graphs depict the mean squared error (MSE) per pattern for the best model in the population as a function of the number of generations. Training was performed over 138 patterns, with one additional pattern held out for testing in cross validation. The sufficiency of the number of generations is indicated as the learning curves approach an asymptote.

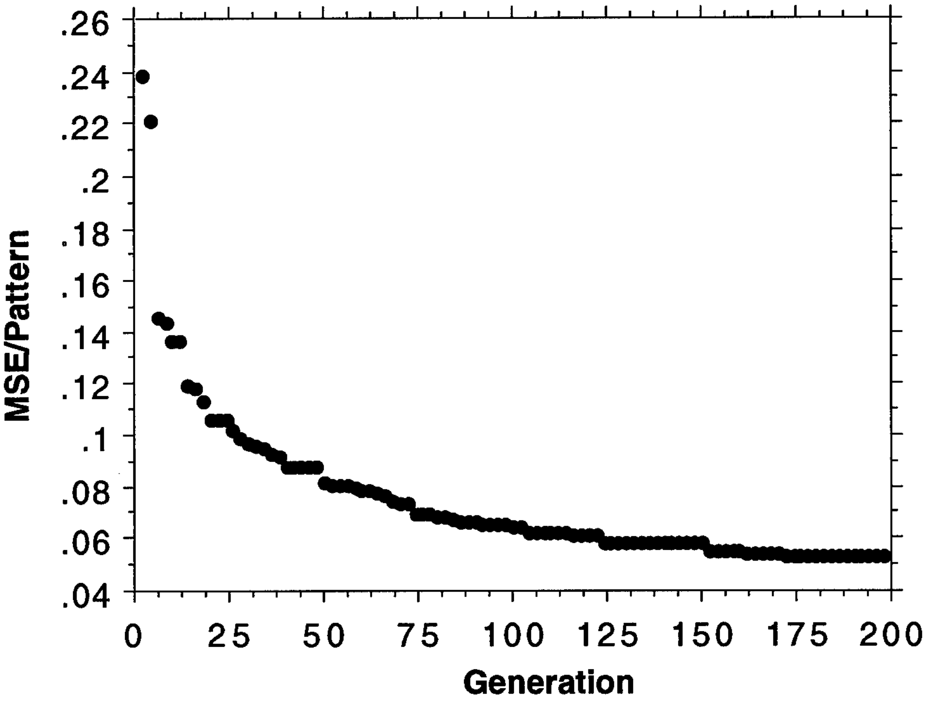


Fig. 7a

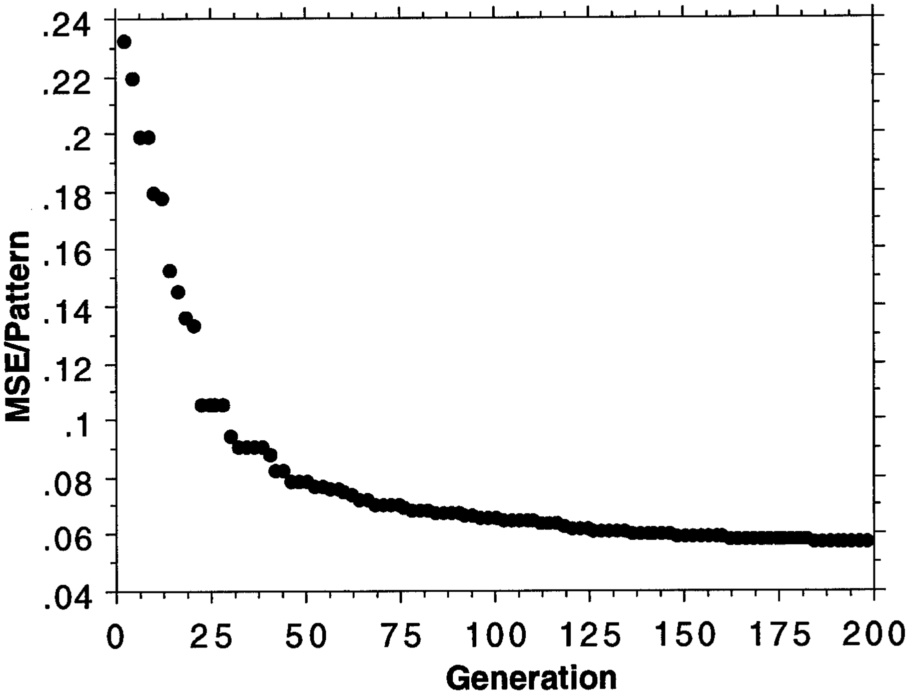

.24
.22
.2
.18
.16
.14
.12
.1
.08
.06
.04
MSE/Pattern
0
25
50
75
100
125
150
175
200
Generation

Figure 7b

Figure 8. Typical ROC curves (raw data) generated for (a) the ANN and (b) the linear classifier in one complete cross validation where each of 139 patterns was classified in turn, based on training over the remaining 138 patterns. Each point represents the probability of detection, P(D), and probability of false positive, P(FP), that is attained as the threshold for classifying a result as malignant is increased systematically over [0,1].

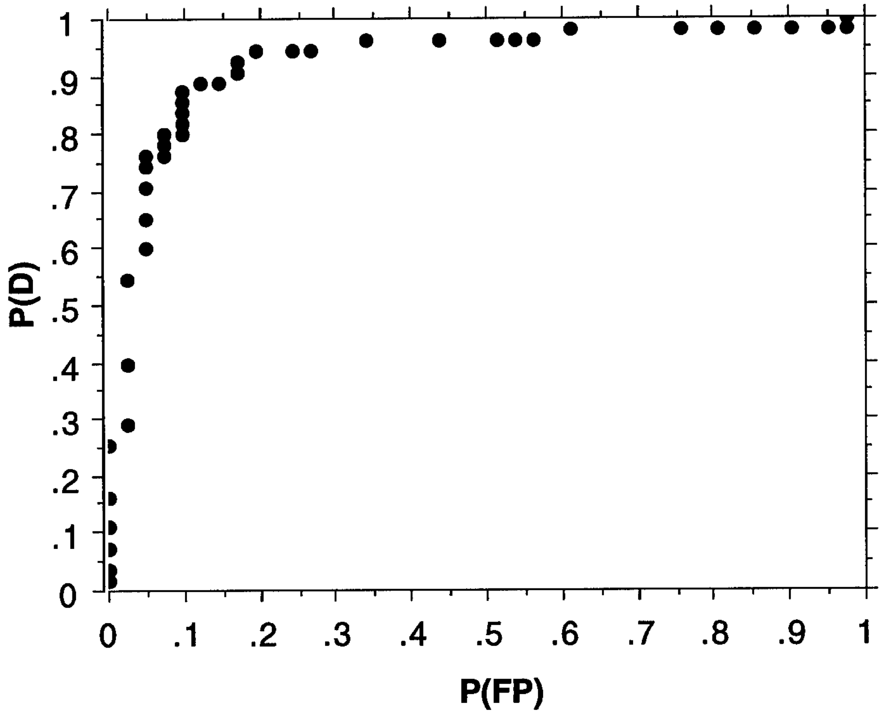


Figure 8a

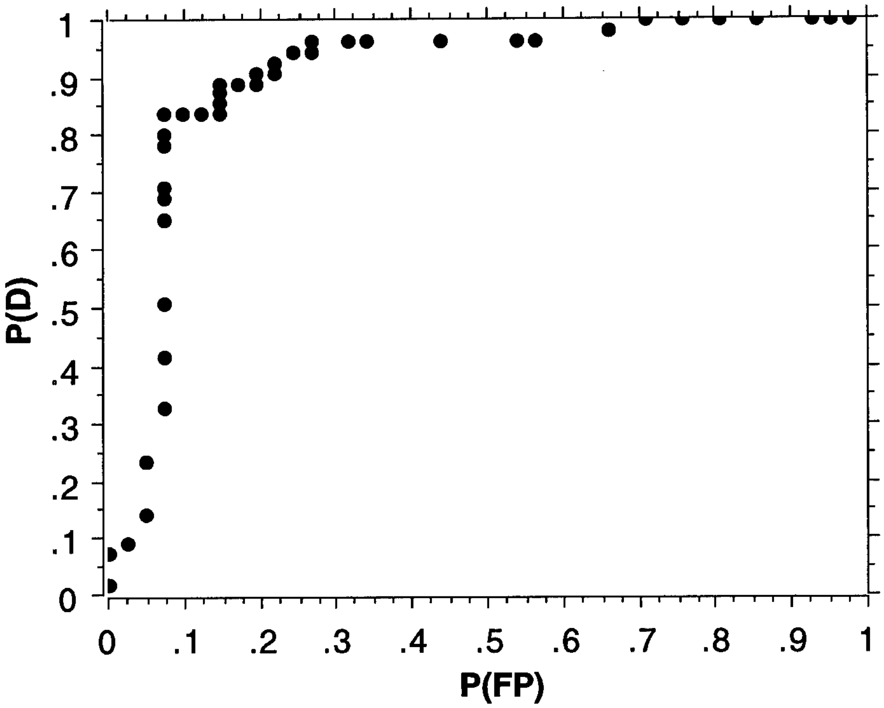


Figure 8b

Figure 9. Typical ROC curves (raw data) generated for the ANN classifier in one complete cross validation where each of 158 patterns was classified in turn, based on training over the remaining 157 patterns. Each point represents the probability of detection, P(D), and probability of false positive, P(FP), that is attained as the threshold for classifying a result as malignant is increased systematically over [0,1]. Using a polynomial spline, the area under the ROC curve depicted here is estimated at 0.9147, slightly below the mean of 0.9196 observed across all 16 trials.

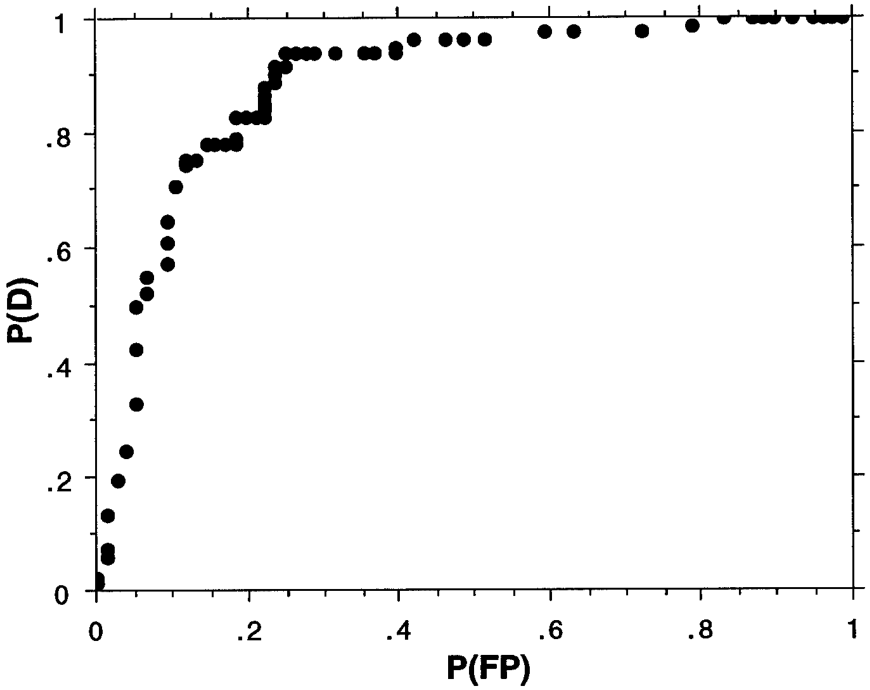


Figure 10. The mean and one standard error bars on the output of the best evolved neural networks across all 158 classifications in the 16th trial. There is a good separation between the two output classes, indicating that the neural classifiers are capable of discriminating between the benign and malignant cases.

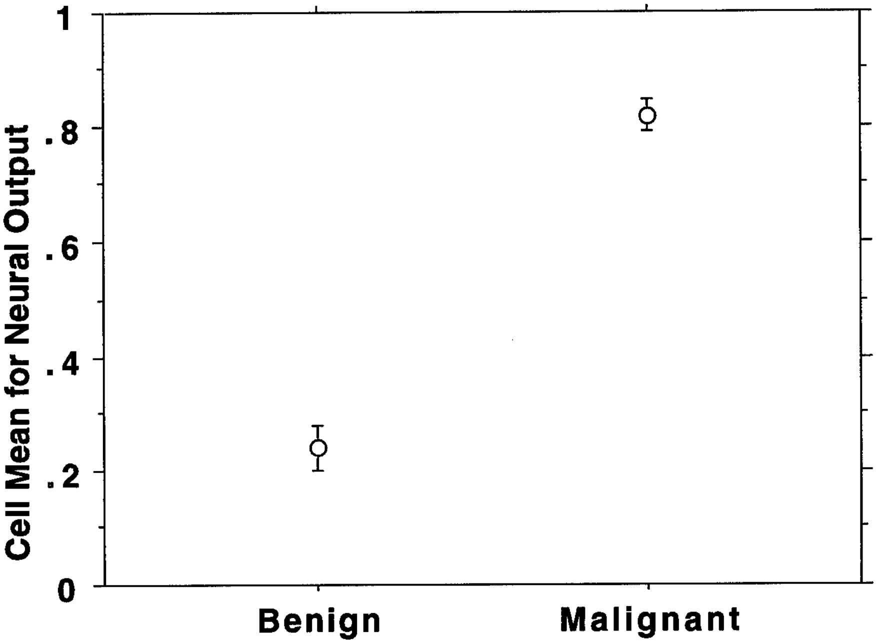
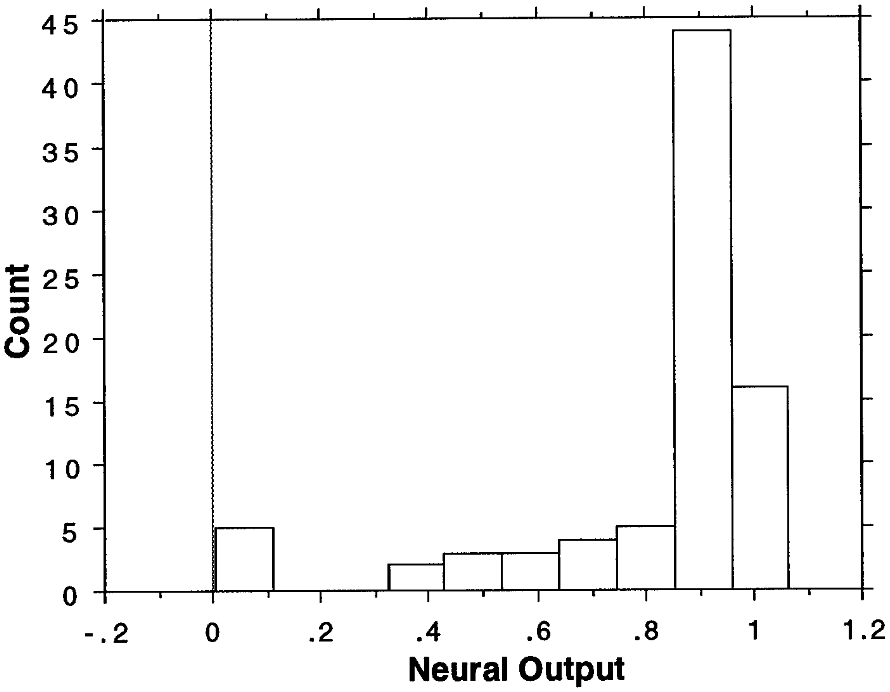
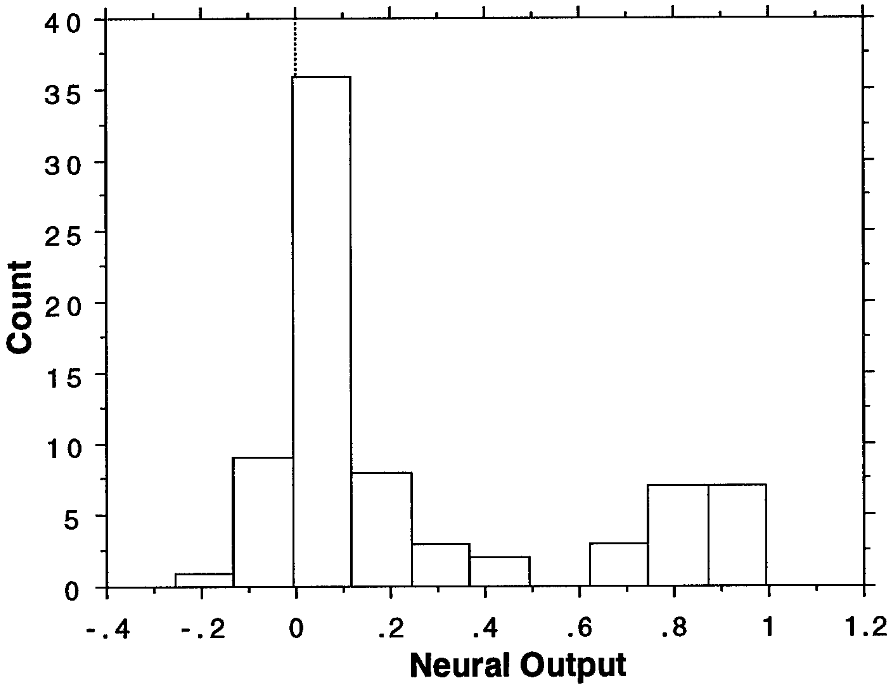


Figure 11. The histograms for the neural output for (a) malignant cases, and (b) benign cases. The distribution of outputs indicates that the two classes are separated well by the evolved neural networks. Malignant cases generally induce large output values, while benign cases generally induce small output values.



(a)



(b)

Figure 12. The ROC curve (raw data) generated for the ANN classifier in one complete cross validation where each of 216 patterns (both suspicious masses and microcalcifications, 111 malignant) was classified in turn, based on training over the remaining 215 patterns. Each point represents the probability of detection, P(D), and probability of false positive, P(FP), that is attained as the threshold for classifying a result as malignant is increased systematically over [0,1]. Using a polynomial spline, the area under the ROC curve depicted here is estimated at 0.8464, lower than could be achieved when focusing solely on suspicious masses.

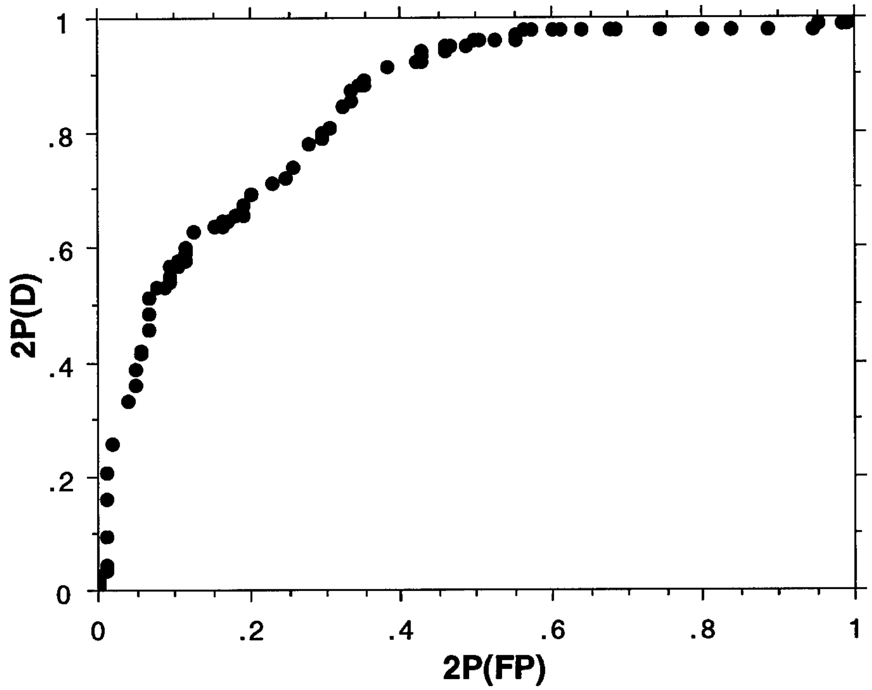
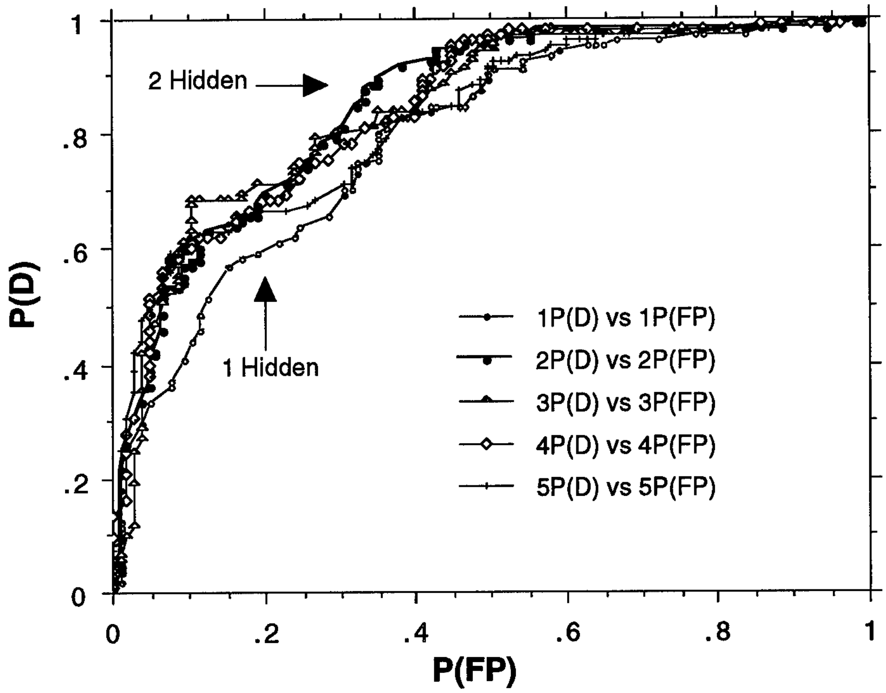


Figure 13. Comparing the ROC curves for neural networks with hidden nodes varying in number from 1 to 5. The network with only a single hidden node is essentially a linear classifier, whereas the other networks can perform nonlinear discrimination. The curves are depicted as connected graphs for ease of comparison. The ROC curve for the linear classifier provides the lowest area under the curve. Including the capability for nonlinear discrimination (i.e., using two hidden nodes) improves performance; however, including additional degrees of freedom in the from of more hidden nodes does not improve on that performance attained by using only two hidden nodes. This provides some justification that the two-hidden node architecture is a reasonable choice for the available data.


P(D)
1
.8
.6
.4
.2
0
0
.2
.4
.6
.8
1
P(FP)
2 Hidden
1 Hidden
1P(D) vs 1P(FP)
2P(D) vs 2P(FP)
3P(D) vs 3P(FP)
4P(D) vs 4P(FP)
5P(D) vs 5P(FP)

Appendices

Bibliography of publications arising from funded effort

1. D.B. Fogel, E.C. Wasson, E.M. Boughton, V.W. Porto, and J.W. Shively (1997) "Initial results of training neural networks to detect breast cancer using evolutionary programming," *Control and Cybernetics*, in press.

2. D.B. Fogel, E.C. Wasson, E.M. Boughton, and V.W. Porto (1997) "A step toward computer-assisted mammography using evolutionary programming and neural networks," *Cancer Letters*, in press.

3. D.B. Fogel, E.C. Wasson, E.M. Boughton, and V.W. Porto (1997) "Evolving linear and neural models for classifying breast masses," *IEEE Trans. Med. Imaging*, in review.

4. D.B. Fogel, E.C. Wasson, and E.M. Boughton (1997) "Using neural networks in diagnosing breast cancer," Extended Abstract, USAMRMC Breast Cancer Research Program: An Era of Hope, Washington DC, October, in press.

List of personnel receiving pay

Natural Selection, Inc.
Dr. David B. Fogel
Vincent W. Porto
Jamen Shivley
Eva Fogel

Consultants
Eugene C. Wasson, M.D.
Edward M. Boughton

Papers Resulting from Funded Research

Paper in press at *Control and Cybernetics,* 1997.

Initial Results of Training Neural Networks to Detect Breast Cancer using Evolutionary Programming

David B. Fogel
Natural Selection, Inc.
3333 N. Torrey Pines Ct., Suite 200
La Jolla, CA 92037
dfogel@natural-selection.com

Eugene C. Wasson
Maui Memorial Hospital
221 Mahalani
Wailuku, HI 96793
wasson@maui.com

Edward M. Boughton
Hawaii Industrial Laboratory, Inc.
P.O. Box 1275
Wailuku, HI 96793
boughton@maui.com

Vincent W. Porto
Jamen W. Shively
Natural Selection, Inc.
3333 N. Torrey Pines Ct., Suite 200
La Jolla, CA 92037
bporto@natural-selection.com
jshively@natural-selection.com

Abstract

Artificial neural networks are applied to the problem of detecting breast cancer from radiographic features and patient age. Evolutionary programming is used to train neural networks based on sigmoid or Gaussian kernel functions. Preliminary results on 96 biopsy-proven cases (62 malignant, 34 benign) indicate that a reasonable probability of detecting malignancies can be achieved using simple neural architectures. The features appear to be more amenable to discrimination by partitioning functions than to clustering functions, although final analysis remains for larger sample sizes.

Introduction

Carcinoma of the breast is second only to lung cancer as a tumor-related cause of death in women. There are now more than 180,000 new cases and 45,000 deaths annually in the United States alone. It begins as a focal curable disease, but it is usually not identifiable by palpation at this stage, and mammography remains the mainstay in effective screening. It has been estimated that the mortality from breast carcinoma could be decreased by as much as one-third if all women in the appropriate age groups were regularly screened.

Computer technology offers many potential benefits to the radiologist, including computer-aided diagnosis. There is currently considerable intra- and inter-observer disagreement or inconsistencies in mammographic interpretation. This has led to an interest in the possibility of utilizing computerized pattern recognition algorithms, such as artificial neural networks (ANNs), to assist in the decision-making required in the assessment of mammograms. ANNs have been demonstrated to be useful in many engineering pattern recognition applications and these techniques hold promise for improving the accuracy of determining those patients where further assessment and possible biopsy is indicated. Furthermore, there should also be an eventual cost savings when a reliable automated screening system can be developed. The successful development of a neural network that is capable of reliably assessing the potential for the existence of breast carcinoma based on radiographic features of mammograms would make the radiologist both more efficient and more effective.

ANNs are models based on the neuronal structure of natural organisms (Haykin, 1994). They are stimulus-response transfer functions that accept some input and yield some output. They are typically used to learn an input-output mapping over a set of examples. For example, as will be described here, the input can be radiographic features from mammograms, with the output being a decision regarding the likelihood of a malignancy. Hornik et al. (1989) and Poggio and Girosi (1990) have proved that neural networks with sigmoid or Gaussian basis functions in a single hidden layer can in principle generate any measurable mapping, indicating the versatility of these functions.

Given a network architecture (i.e., type of network, the number of nodes in each layer, the connections between the nodes, and so forth), and a training set of input patterns, the collection of variable weights determines the output of the network to each presented pattern. The error between the actual output of the network and the desired target output defines a response surface over a hyperspace having a dimension equal to the number of weights. A commonly employed method for finding weight sets in such applications is error *back propagation*, which is essentially a gradient method. As such, it is subject to entrapment in locally optimal solutions, and the resulting weight sets are often unsuitable for practical applications. Numerical optimization techniques that do not suffer from such entrapment can be used to advantage in these cases.

Evolutionary algorithms offer one such technique. In these stochastic optimization methods, a population of candidate solutions is maintained, and random variation (mutation and/or recombination) and selection are imposed on the population to guide it to appropriate regions of the hyperspace. The use of random variation to bias the search avoids entrapment in local optima, and there are several mathematical proofs that variations of these procedures provide asymptotic global convergence, rather than merely local convergence (Fogel, 1994; Rudolph, 1994; Bäck, 1996). Moreover, there is empirical evidence that the methods are robust to many pathologies in possible response surfaces, including multiple minima or maxima, constraints, disjoint feasible regions, and random perturbations (Schwefel, 1995; Fogel, 1995; Michalewicz, 1996; and others).

There have been many efforts to train neural networks using evolutionary algorithms (Fogel et al., 1990; Angeline et al., 1994; McDonnell and Waagen, 1994; Yao and Liu, 1996; and many others). This paper describes the results of preliminary efforts to use evolutionary programming to train simple ANNs to respond to a set of radiographic features from film screen mammograms, along with the patient's reported age, to make a determination regarding the presence or absence of a malignant condition. It begins with a brief review of selected efforts to use neural networks in breast cancer detection, before describing the current methods and results.

Background

Neural networks have been receiving recent attention in medical diagnostics (Brotherton and Simpson, 1995; Rizki et al., 1995; and others). With regard to detecting breast cancer, efforts have been directed at classifying histologic data from cells removed by fine needle aspiration (Wolberg et al., 1994, 1995) and radiographic features from film screen mammography (Kocur et al., 1995; and others). Three of these efforts are reviewed here.

The investigation of Wu et al. (1993) used 43 preselected features related to density, microcalcification, parenchymal distortion, skin thickening, correlation with clinical findings, and so forth. Data was taken from 133 textbook cases in Tabar and Dean (1985). For each mammogram, each of the selected features was rated by an experienced mammographer on a scale of 0-10, and this served as the vector input to a multilayer perceptron neural network (i.e., feedforward and fully connected). The network possessed 10 hidden units and a single output unit which was trained to yield a value of 0.0 for a benign case and 1.0 for a malignancy. Training was accomplished using back propagation. The results of this preliminary study and other described experiments indicated the suitability of this approach. By pruning the feature set to a more reasonable, smaller collection, the neural network was able to statistically outperform an attending radiologist and residents in assessing patterns of mammographic image features that are associated with benign and malignant lesions. There was no statistically significant difference between the performance of the network and the experienced mammographer used to rate each of the image features.

Floyd et al. (1994) used back propagation on multilayer perceptrons to predict breast cancer from mammographic findings from patients who were scheduled for biopsy. They used only eight input parameters (mass size, mass margin, asymmetric density, architectural distortion, calcification number, calcification morphology, calcification density, and calcification distribution) and each of these was parameterized less subjectively than it appears in Wu et al. (1993). There were 260 cases used for training and testing. Data was not separated into complementary training and testing sets; all of the exemplars were processed using a jackknife statistical procedure. After significant training, the results indicated that if a threshold value of 0.1 were used (output on a scale from [-1,1]), 38 out of 168 benign cases and all 92 malignancies would be identified. The authors compared this performance to that of radiologists and suggested that these results were statistically significantly better than radiologists at a $P < 0.08$ level. Although their results do appear fairly impressive with regard to detecting malignancy, the number of false alarms is somewhat high (23%), and the statistical validity of the hypothesis test carried out can be questioned because the threshold of 0.1 was chosen after the authors reviewed the data and statistics were compiled on those same data. Thus the data did not reflect a random sample, but rather a biased sample. New data would have to be tested at the threshold value of 0.1 to ensure a sound statistical procedure.

Wilding et al. (1994) used back propagation on multilayer perceptrons to assess both breast and ovarian cancer. Their procedure was similar to Wu et al. (1993) and Floyd et al. (1994), except that their input parameters consisted mainly of objective blood specimens and analyses (maximum of 10 total input parameters) from 104 patients. Unfortunately, Wilding et al. (1994) reported that the neural network was able to "provide little improvement on the sensitivity of testing comparing to the use of [the tumor marker] CA 15-3 only. Furthermore, it would appear that none of the networks appear to identify any worthwhile parameters or operating conditions with clinical utility."

Wu et al. (1993), Floyd et al. (1994), and Wilding et al. (1994) each used back propagation to determine the weights of their neural networks. But networks trained by gradient methods may require many more hidden nodes to train to a tolerable level of error than are actually required because the method may converge at suboptimal weight sets.

Adding more weights (i.e., degrees of freedom) can help overcome local optima and offer the possibility for suitable training, but overparameterized networks may not generalize well on new data. These concerns were specifically discussed in Floyd et al. (1994) and Wilding et al. (1994). The most effective methods employed in these investigations for limiting the number of nodes and network parameters were based on sensitivity analysis and ad hoc pruning. Sensitivity analysis is problematic on nonlinear transfer functions (such as neural networks) and ad hoc pruning can be largely unproductive. Despite directly mentioning concerns about overfitting their data, Floyd et al. (1994) found that their best performance occurred when using 177 weights (16 hidden nodes), but they used only 260 samples. Wilding et al. (1994) used networks with as few as 38 weights and as many as 132 weights, and despite having 104 samples were still unable to generate satisfactory performance. Even if the blood statistics that were being used were not particularly relevant to the classification task at hand, the failure to find suitable networks with more parameters than data indicates the limitations of the training method and suggests alternative methods for optimizing classification networks, such as evolutionary computation.

Method

Data for the current effort were collected by assessing film screen mammograms in light of a set of radiographic features as determined by the domain expert (Wasson). The features selected paralleled those of Floyd et al. (1994) with some important modifications. Under the system of Floyd et al. (1994), certain features were described as lying on a continuum when it appeared more useful to rate these features independently. For example, Floyd et al. (1994) rated mass margin with six categories: (1) no mass (value 0.0), (2) well circumscribed (value 0.2), (3) microlobulated (value 0.4), (4) obscured (value 0.6), (5) indistinct (value 0.8), and (6) and spiculated (1.0). In contrast, the current parameterization rated the five categories of masses (see (2)-(6) in Table 1) in four levels [0,1,2,3] as none, low, medium, and high. The complete set of radiographic features used appears in Table 1. In addition, the age of the patient was considered leading to a total of 13 input features. These features were assessed in 96 cases all of which subsequently had open surgical biopsy of the area of concern, with the associated pathology indicating whether or not a malignant condition had been found. In all, 62 cases were associated with a biopsy-proven malignancy, while 34 cases were indicated to be negative by biopsy (although the possibility remains that such an indication may be in error).

These data were processed using two forms of neural networks: (1) multilayer perceptrons and (2) receptive fields. Each network architecture was restricted to two hidden nodes, with a linear output node, resulting in 31 adjustable weights (see Figure 1). The perceptron network used a sigmoid filter on each hidden node of $f(\beta) = (1 + \exp(-\beta))^{-1}$, where β is the sum of a bias term and the dot product of the input feature vector and the associated weight vector. The receptive field network used a Gaussian filter on each hidden node of $f(\beta) = (2\pi)^{-0.5}\exp(-\beta^2)$. Evolutionary programming was used to train the networks in a leave-one-out cross validation procedure.

Specifically, for each complete cross validation where each sample pattern in turn was held out for testing then replaced in a series of 96 separate training procedures, a population of 250 networks of the chosen architecture were selected at random by sampling weight values from a uniform random variable distributed over [-0.5,0.5]. Each weight set (i.e. candidate solution) also had an associated self-adaptive mutational vector used to determine the random variation imposed during the generation of offspring networks (described below). Each of the self-adaptive parameters was initialized to a value of 0.01. Each weight set was evaluated based on how well the network classified the 95 remaining available training patterns (with one "left out" for testing), where a malignant condition was assigned a target value of 1.0 and a benign conditions was assigned a target of 0.0. The performance of each network was determined as the sum of the squared error between the output and the target value taken over the 95 available patterns.

After evaluating all existing (parent) networks, the 250 weight sets were used to generate 250 offspring weight sets (one offspring per parent). This was accomplished in a two-step procedure. For each parent, the self-adaptive parameters were updated as:

$$\sigma'_i = \sigma_i \exp\left(\tau N(0,1) + \tau' N_i(0,1)\right) \qquad (1)$$

where t = $\frac{1}{\sqrt{2n}}$, t' = $\frac{1}{\sqrt{2\sqrt{n}}}$, $N(0,1)$ is a standard normal random variable sampled once for all $n = 33$ parameters of the vector σ, and $N_i(0,1)$ is a standard normal random variable sampled anew for each parameter. The settings for τ and τ' have been recommended as robust in Bäck and Schwefel (1993). These updated self-adaptive parameters were then used to generate new weight values for the offspring according to the rule:

$$x'_i = x_i + \sigma'_i C \qquad (2)$$

where C is a standard Cauchy random variable

$$f(y) = \frac{1}{\pi\left(1 + y^2\right)}, -\infty < y < \infty \qquad (3)$$

(determined as the ratio of two independent standard Gaussian random variables). Traditional methods in evolutionary programming and evolution strategies have relied on Gaussian mutation, however recent research in Yao and Liu (1996), Saravanan and Fogel (1997), and others, have suggested a possible benefit to using Cauchy variation because it has a greater probability to generate longer jumps than the Gaussian. This offers a greater chance of escaping local optima on a error surface at the expense of poorer fine tuning. Initial observations with both Gaussian and Cauchy mutations on the existing data appeared to favor the Cauchy distribution, however a more careful analysis remains for future study. All of the offspring weight sets were evaluated in the same manner as their parents.

Selection was applied to eliminate half of the total parent and offspring weight sets based on their observed error performance. Following typical methods in evolutionary programming (Fogel 1995), a pairwise tournament was conducted where each candidate weight set was compared against a random sample from the population. The sample size was chosen as 10 (a greater sample size indicates more stringent selection pressure). For each of the 10 comparisons, if the weight set had an associated classification error score that was lower than the randomly sampled opponent it received a "win." After all weight sets had participated in this tournament, those that received the greatest number of wins were retained as parents of the next generation.

This process was iterated for 100 generations, whereupon the best available network as measured by the training performance was used to classify the held out input feature vector. The result of this classification was recorded (i.e., the output value of the network and the associated target value) and the process was restarted by replacing the held out vector and removing the next vector in succession until all 96 patterns had been classified.

Each complete series of cross validation was repeated 10 times for both the multilayer perceptron and receptive field networks.

Results

A typical rate of optimization in each training run is shown in Figure 2. The overall training error often fell as a nearly linear function of the number of generations without saturation. This suggests that further training time might be warranted.

The probability of detection, P(D), and false alarm, P(FA), vary as a function of the discrimination threshold applied to the output of the networks. As the threshold value is lower, the network can correctly identify a greater number of cancers, but this comes at the expense of a higher false alarm rate. Conversely, the false alarm rate can be lowered by raising the threshold value, but this in turn decreases the sensitivity of the procedure.

The effectiveness of the classification procedures can be assessed using receiver operating characteristic (ROC) analysis, where the probability of detecting a malignancy is traded off as a function of the likelihood of a false positive result. Typical ROC curves for the multilayer perceptron and receptive field networks are shown in Figure 3. The area under the curve provides a useful measure for comparison.

To compare the effectiveness of the multilayer perceptron architecture with the receptive field, the area under the ROC curve for each of the 10 trials of cross validation with each method was estimated. This was accomplished by performing a polynomial regression of at least third order to the available samples of fraction of false alarms versus fraction of detections in each ROC curve. Regression models were determined by choosing the lowest order polynomial that provided (1) an R^2 value of at least 0.99, and (2) was non-decreasing over false alarm rates from zero to one (see Figure 4). The models were integrated over the range [0,1] to computer the desired areas. The mean area under the ROC curve (and standard deviation) for the perceptron and receptive field networks, respectively, were 0.787611±0.022346 and 0.739060±0.035574. Under a two-sample mean *t*-test (which assumes populations of normally distributed values), these data indicate statistically significant evidence in favor of the perceptron (sigmoid) networks ($P < 0.01$).

Conclusions

Under the assumptions of normally distributed integrals of the ROC curves, the data suggest that partitioning functions (as offered by sigmoid filters) may be more useful than clustering functions (as offered by Gaussian filters) for classifying the radiographic features and patient age as being indicative of a breast malignancy. The longer-term relevance of this result is yet unclear due at least to (1) the relatively small sample size, and (2) the constraint that all patterns were derived from mammograms that presented sufficient radiographic findings to suggest biopsy. Current efforts are directed to obtaining a larger sample.

Comparisons of the overall performance offered here with the results offered in Floyd et al. (1994) must be made with caution. The composition of the 260 samples in Floyd et al. (1994) was 64.6% benign cases, with only 35.4% malignancies. In contrast, the current data set offered almost the obverse conditions. Further, the demographics between the studies were different. The data in Floyd et al. (1994) were derived from examinations at Duke University Medical Center, whereas the data for the current study were collected from radiology centers on the island of Maui, which can be expected to provide a more diverse racial mix (24% part-Hawaiian, 22% Caucasian, 17% Japanese, and so forth). This greater diversity might be expected to pose a more significant challenge for a classification algorithm.

In separate analysis, evolutionary programming was used to train a two-hidden node perceptron over the entire 96 available patterns. Using an output threshold of 0.5 (i.e., greater than 0.5 indicates a diagnosis of a malignancy), it was possible, after more than 2000 generations, to find a weight vector that misclassified only 3 of the 96 patterns (i.e., it was in error on two malignancies and one benign case). Yet when this same architecture was used in the cross validation trials, this degree of overall performance was not attained. This suggests that (1) further training in the cross validation trials may be useful, and/or (2) the current neural architecture overfits the available data, in which case future analysis on a larger collection of samples should yield a closer correspondence between the error rates when training on all available data and when training/testing in cross validation. Other possibilities for improving the discrimination performance of the evolved networks include imposing small amounts of random noise on the input patterns to increase the possible generalizability (i.e., essentially creating a larger sample size) and reducing the degrees of freedom of the neural networks by limiting the input parameters to a subset of the current assortment.

Acknowledgments

This effort is sponsored by the Department of the Army, Grant No. DAMD17-96-1-6312. The content of the information presented does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Thanks are owed to M.A. Williams for facilitating the use of the SP-2 at the Maui High Performance Computing Center. Research sponsored in part by the Phillips Laboratory, Air Force Materiel Command, USAF, under cooperative agreement number F29601-93-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Phillips Laboratory or the U.S. Government.

References

P.J. Angeline, G.M. Saunders, and J.B. Pollack (1994) "An evolutionary algorithm that constructs recurrent neural networks," *IEEE Trans. Neural Networks,* **5**:1, 54-65.

T. Bäck (1996) *Evolutionary Algorithms in Theory and Practice*, Oxford, NY.

T. Bäck and H.-P. Schwefel (1993) "An overview of evolutionary algorithms in parameter optimization," *Evolutionary Computation,* **1**:1, 1-24.

T.W. Brotherton and P.K. Simpson (1993) "Dynamic feature set training of neural nets for classification," *Evolutionary Programming IV: Proceedings of the Fourth Annual Conference on Evolutionary Programming,* J.R. McDonnell, R.G. Reynolds, and D.B. Fogel (eds.), MIT Press, Cambridge, MA, pp. 83-94.

D.B. Fogel (1994) "Asymptotic convergence properties of genetic algorithms and evolutionary programming: Analysis and experiments," *Cybernetics and Systems,* **25**:3, 389-407.

D.B. Fogel (1995) *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, IEEE Press, NY.

D.B. Fogel, L.J. Fogel, and V.W. Porto (1990) "Evolving neural networks," *Biological Cybernetics*, **63**:6, 487-493.

C.E. Floyd, J.Y. Lo, A.J. Yun, D.C. Sullivan, and P.J. Kornguth (1994) "Prediction of breast cancer malignancy using an artificial neural network," *Cancer*, **74**, 2944-2998.

S. Haykin (1994) *Neural Networks: A Comprehensive Foundation*, MacMillan, NY.

K. Hornik, M. Stinchcombe, and H. White (1989) "Multilayer feedforward networks are universal approximators," *Neural Networks*, **2**, 359-366.

C.M. Kocur, S.K. Rogers, K.W. Bauer, and J.M. Steppe (1995) "Neural network feature selection for breast cancer diagnosis," *Applications and Science of Artificial Neural Networks*, S.K. Rogers and D.W. Ruck (eds.), Proc. SPIE 2492, 905-918.

J.R. McDonnell and D. Waagen (1994) "Evolving recurrent perceptrons for time-series modeling," *IEEE Trans. Neural Networks*, **5**:1, 24-38.

Z. Michalewicz (1996) *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd ed., Springer, Berlin.

T. Poggio and F. Girosi (1990) "Networks for approximation and learning," *Proc. of the IEEE*, **78**:9, 1481-1497.

G. Rudolph (1994) "Convergence analysis of canonical genetic algorithms," *IEEE Trans. Neural Networks*, **5**:1, 96-101.

N. Saravanan and D.B. Fogel (1997) "Multi-operator evolutionary programming," *Evolutionary Programming VI: Proceedings of the Sixth Annual Conference on Evolutionary Programming*, P.J. Angeline, R.C. Eberhart, R.G. Reynolds, and J.R. McDonnell (eds.), Springer, Berlin, in press.

H.-P. Schwefel (1995) *Evolution and Optimum Seeking*, John Wiley, NY.

L. Tabar and P.B. Dean (1985) *Teaching Atlas of Mammography*, 2nd ed., Thieme-Stratton, NY.

P. Wilding, M.A. Morgan, A.E. Grygotis, M.A. Shoffner, and E.F. Rosato (1994) "Application of backpropagation neural networks to diagnosis of breast and ovarian cancer," *Cancer Letters*, **77**, 145-153.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian (1995) "Computerized breast cancer diagnosis and prognosis from fine-needle aspirates," *Arch. Surg.*, 130, 511-516.

W.H. Wolberg, W.N. Street, and O.L. Mangasarian (1994) "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer Letters*, **77**, 163-171.

Y.Z. Wu, M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schmidt, and C.E. Metz (1993) "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, **187**:1, 81-87.

X. Yao and Y. Liu (1996) "Evolving artificial neural networks through evolutionary programming," *Evolutionary Programming V: Proceedings of the Fifth Annual Conference on Evolutionary Programming*, L.J. Fogel, P.J. Angeline, and T. Bäck (eds.), MIT Press, Cambridge, MA, pp. 257-266.

Table 1. The features and rating system used for assessing mammograms in the current study. Assessment was made by the domain expert (Wasson).

1. Mass size: either zero or in mm.
2. Mass margin: (each subparameter rated as none (0), low (1), medium (2), or high (3))
    - (a) Well circumscribed
    - (b) Microlobulated
    - (c) Obscured
    - (d) Indistinct
    - (e) Spiculated
3. Architectural distortion: none or distortion
4. Calcification number: none (0), < 5 (1), 5-10 (2), or > 10 (3).
5. Calcification morphology: none (0), not suspicious (1), moderately suspicious (2), or highly suspicious (3)
6. Calcification density: none (0), dense (1), mixed (2), faint (3)
7. Calcification distribution: none (0), scattered (1), intermediate (2), clustered (3)
8. Asymmetric density: either zero or in mm.

Figure 1. The design used for processing data, both for the multilayer perceptron and receptive field neural architectures. Input data are weighted in connections to the two hidden nodes. Each hidden node passes the sum of a bias term (not shown) and the dot product of the weights and inputs through a nonlinear filter. The filter is $f(ß) = 1/(1+e^{-ß})$ for the multilayer perceptron, and $f(ß) = (2\pi)^{-0.5}e^{-ß^2}$ for the receptive field. The output node is a linear filter which performs the sum of a bias term with the dot product of filtered hidden nodes and their associated weights. There are 31 weights given 13 inputs.

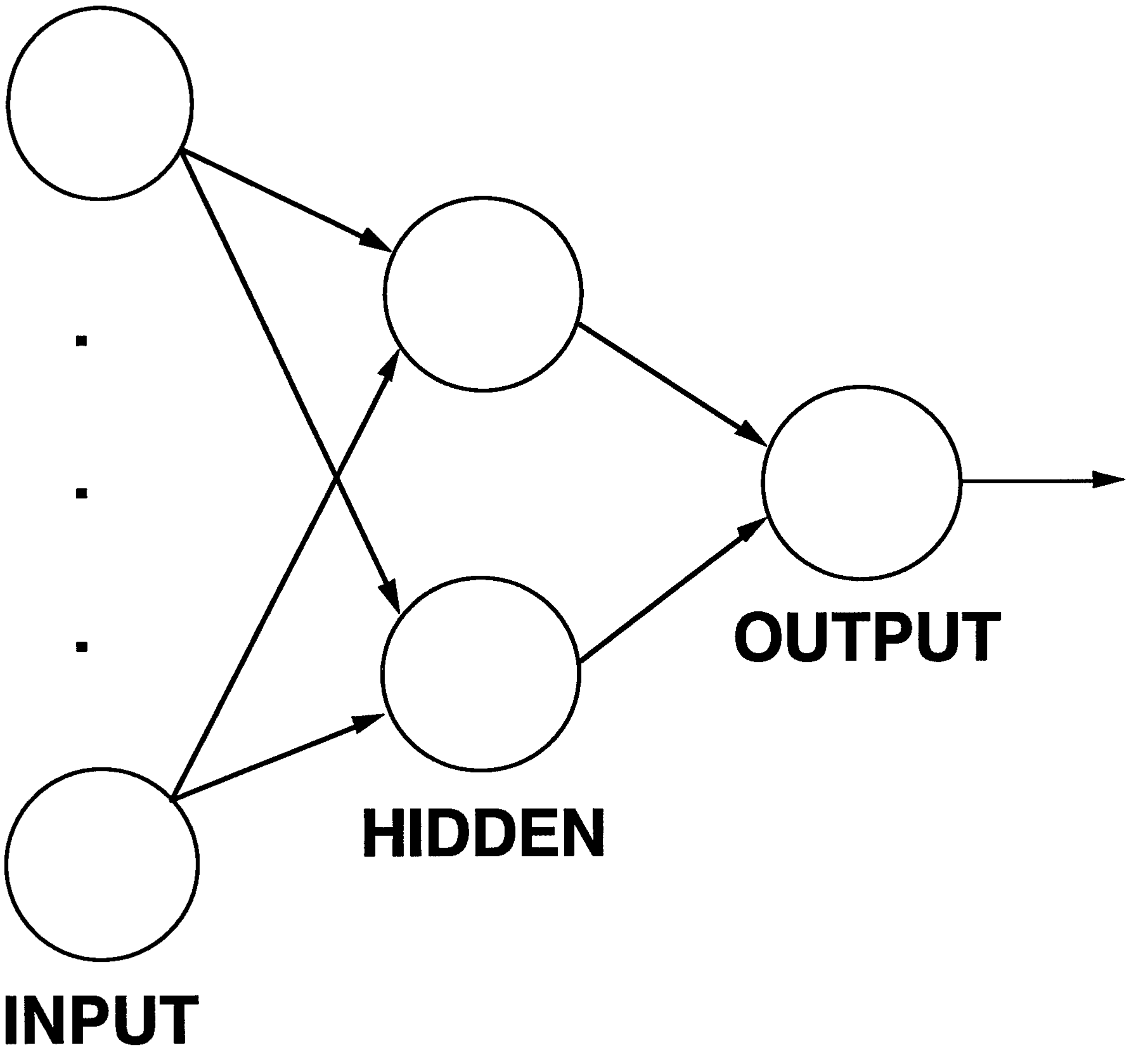


Figure 2. A typical rate of optimization in an evolutionary training of the neural networks on 95 patterns (one held out) over 100 generations. The error of the best member (weight set) in the population is seen to decrease nearly linearly as function of the number of generations. With further training, the observed best error would eventually saturate at an asymptote. The error is taken as the sum of the squared difference between the target value and the realized output from the network generated as a result of each input pattern.

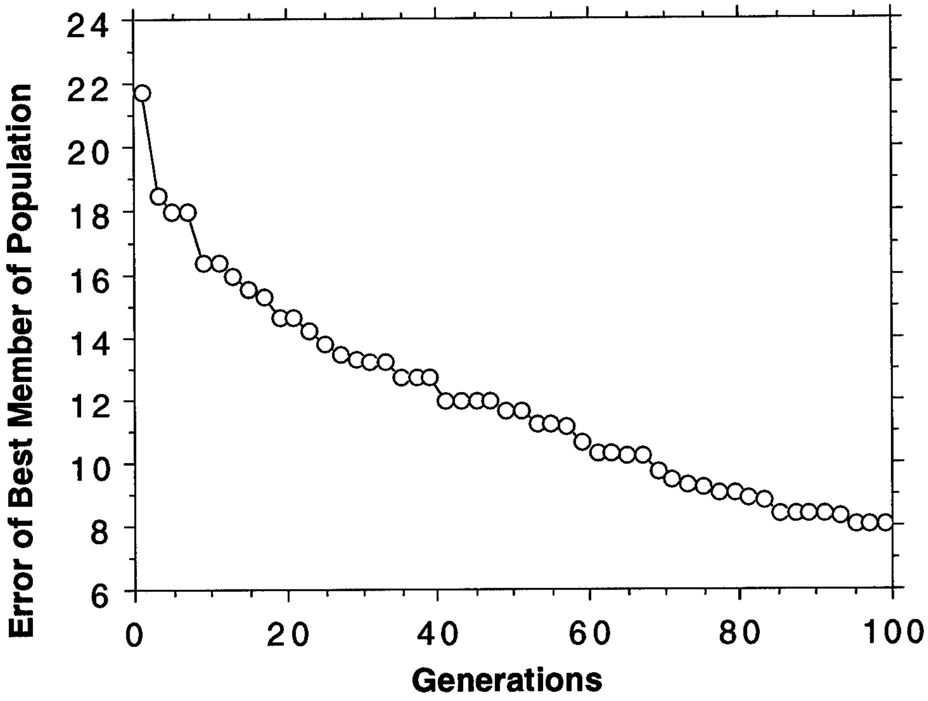
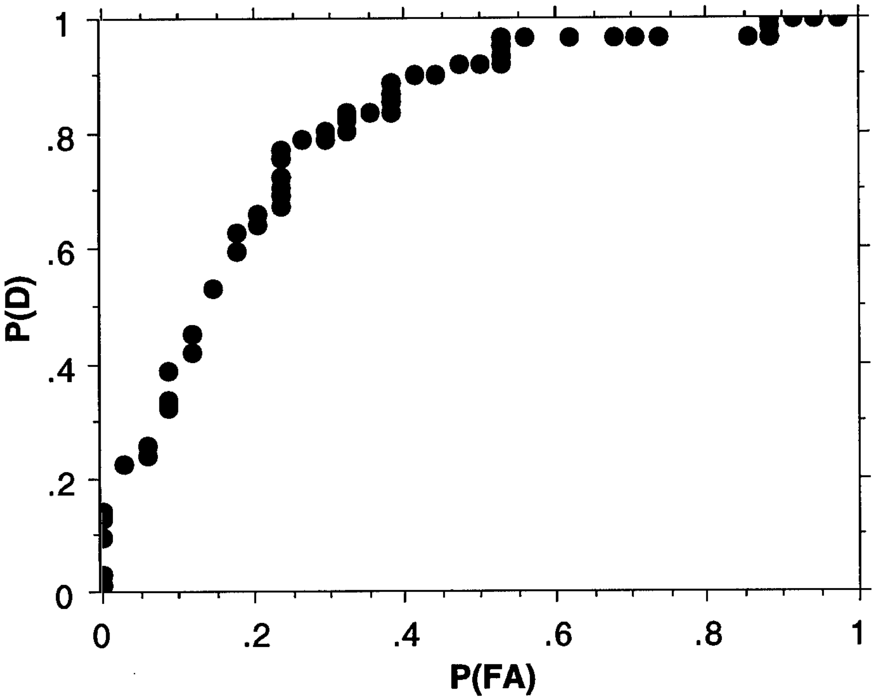
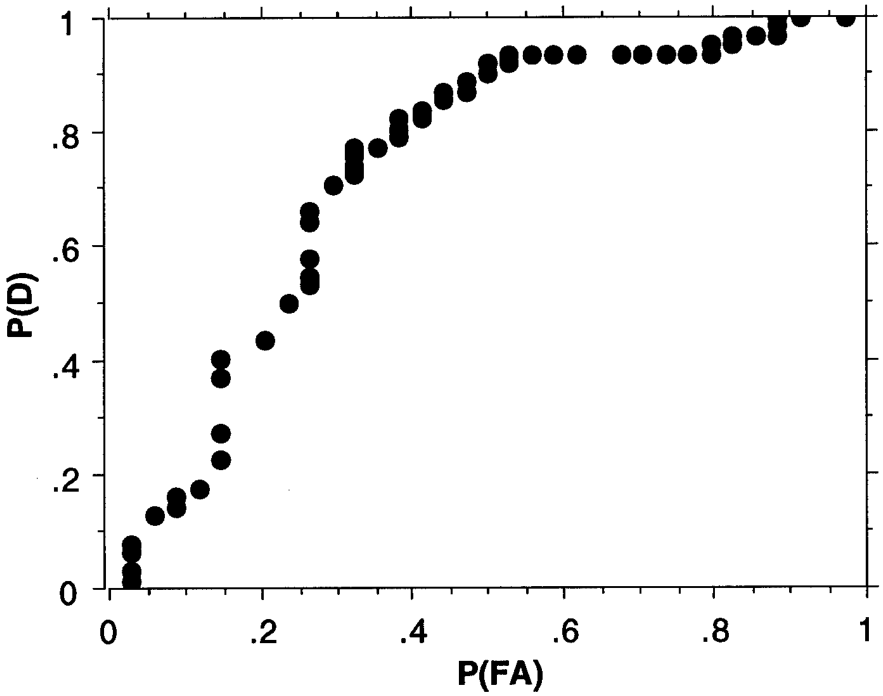


Figure 3. Typical ROC curves for the (a) perceptron and (b) receptive field neural networks. As the probability of a false alarm (i.e., indication of malignancy when none is present) increases, so does the probability of detection.

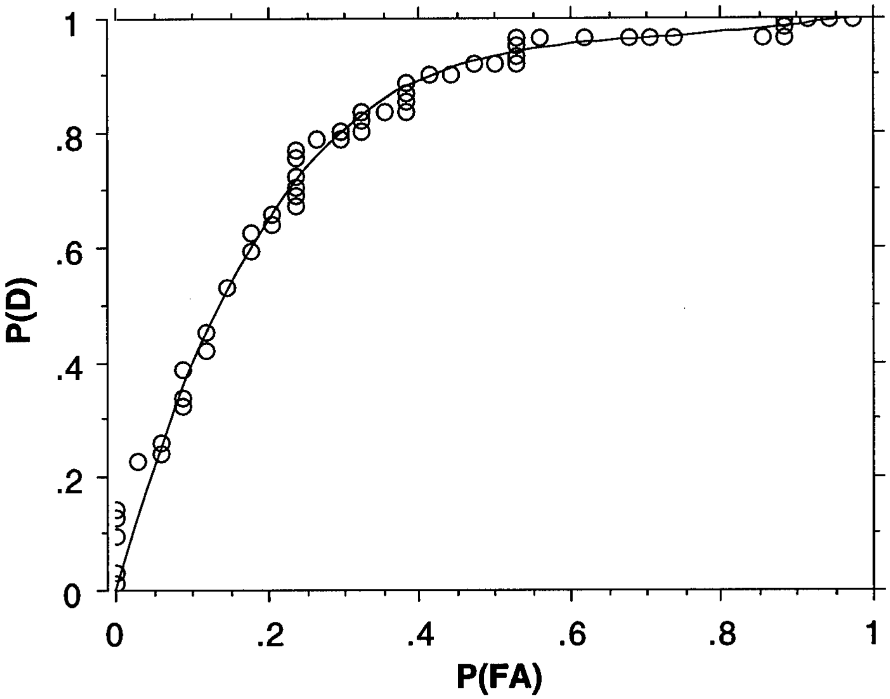


(a)



(b)

Figure 4. An example of using polynomial regression to estimate the ROC curve based on the observed pairs of probabilities for false alarm and detection. The equation shown is P(D) = 4.742 x P(FA) - 8.897 x $P(FA)^2$ + 7.452 x $P(FA)^3$ - 2.295 x $P(FA)^4$. The goodness-of-fit is $R^2 = 0.999$. Note that the regression equation is constrained to pass through the origin. Regressions were conducted for each of the 10 complete cross validation studies with both the perceptron and receptive field networks in order to determine the area under the approximate ROC curve.



Paper in press at *Cancer Letters*, 1997

A Step Toward Computer-Assisted Mammography using Evolutionary Programming and Neural Networks

David B. Fogel
Natural Selection, Inc.
3333 N. Torrey Pines Ct., Suite 200
La Jolla, CA 92037
dfogel@natural-selection.com
(619) 455-6449 tel
(619) 455-1560 fax
(corresponding author)

Eugene C. Wasson
Maui Memorial Hospital
221 Mahalani
Wailuku, HI 96793
wasson@maui.net

Edward M. Boughton
Hawaii Industrial Laboratory, Inc.
P.O. Box 1275
Wailuku, HI 96793
boughton@maui.com

Vincent W. Porto
Natural Selection, Inc.
3333 N. Torrey Pines Ct., Suite 200
La Jolla, CA 92037
bporto@natural-selection.com

Summary

Artificial intelligence techniques can be used to provide a second opinion in medical settings. This may improve the sensitivity and specificity of diagnoses, as well as the cost effectiveness of the physician's effort. In the current study, evolutionary programming is used to train artificial neural networks to detect breast cancer using radiographic features and patient age. Results on 112 suspicious breast masses (63 malignant, 49 benign, biopsy proven) indicate that a significant probability of detecting malignancies can be achieved using simple neural architectures at the risk of a small percentage of false positives.

Keywords: Breast cancer, computer-assisted diagnosis, artificial neural networks, evolutionary computation, evolutionary programming

Introduction

Carcinoma of the breast is second only to lung cancer as a tumor-related cause of death in women. There are now more than 180,000 new cases and 45,000 deaths annually in the United States alone [1]. It begins as a focal curable disease, but it is usually not identifiable by palpation at this stage, and mammography remains the mainstay in effective screening. It has been estimated that the mortality from breast carcinoma could be decreased by as much as 25 percent if all women in the appropriate age groups were regularly screened [2].

Intra- and inter-observer disagreement and inconsistencies in mammographic interpretation [3, 4] have led to an interest in using computerized pattern recognition algorithms, such as artificial neural networks (ANNs) [5], to assist the radiologist in the assessment of mammograms. The "second opinion" offered by a reliable automated system may be useful in reducing false-negative diagnoses [6, 7], and other oversights that may result from poor mammographic image quality, physician fatigue, or alternative sources. ANNs hold promise for improving the accuracy of determining those patients where further assessment and possible biopsy is indicated. Furthermore, a reliable automated screening system could provide immediate results and lower the logistical costs associated with handling mammograms. The eventual cost savings could be passed along to the patient, while simultaneously making the radiologist's time more valuable and improving the quality of patient care.

ANNs are computational models based on the neuronal structure of natural organisms. They are stimulus-response transfer functions that map an input space to a specified output space. They are typically used to generalize such an input-output mapping over a set of specific examples. For example, as will be described here, the input can be radiographic features from mammograms, with the output being an indication of the likelihood of a malignancy.

Given a network architecture (i.e., type of network, the number of nodes in each layer, the weighted connections between the nodes, and so forth), and a training set of input patterns, the collection of variable weights determines the output of the network to each presented pattern. The error between the actual output of the network and the desired target output defines a potentially multimodal response surface over a multidimensional hyperspace (the dimension is equal to the number of weights). A commonly employed method for finding weight sets in such applications is error back propagation, which is essentially a gradient method. As such, it is subject to entrapment in locally optimal solutions, and the resulting weight sets are often unsuitable for practical applications [8]. Numerical optimization techniques that do not suffer from such entrapment can be used to advantage in these cases.

Evolutionary algorithms offer one such technique. In these stochastic optimization methods [9], a population of candidate solutions is maintained, and random variation and selection are imposed on the population to efficiently guide it to appropriate regions of the hyperspace. The use of random mutation avoids entrapment in local optima, and there are several mathematical proofs that variations of these procedures provide asymptotic global convergence, rather than merely local convergence [9, 10]. Moreover, there is empirical evidence that the methods are robust to many difficulties in possible response surfaces, including multiple minima or maxima, constraints, disjoint feasible regions, and random perturbations [11].

Method

For the current investigation, data were collected by assessing film screen mammograms in light of a set of 12 radiographic features as determined by the domain expert (Wasson) (Table I). The features selected paralleled those offered in [12], with some modifications to increase the orthogonality of the features, as well as the inclusion of patient age. These features were assessed in 112 cases of suspicious breast mass, all of which were subsequently examined by open surgical biopsy with the associated pathology indicating whether or not a malignant condition had been found. In all, 63 cases were associated with a biopsy-proven malignancy, while 49 cases were indicated to be negative by biopsy.

These data were processed using a simple feedforward ANN restricted to two hidden sigmoid nodes (following the maxim of parsimony, this being the simplest architecture that can take advantage of the nonlinear properties of the nodes), with a single linear output node, resulting in 33 adjustable weights. Evolutionary programming was used to train the networks in a leave-one-out cross validation procedure. Specifically, for each complete cross validation where each sample pattern was held out for testing and then replaced in a series of 112 separate training procedures, a population of 250 networks of the chosen architecture were initialized at random by sampling weight values from a uniform random variable distributed over [-0.5,0.5]. Each weight set (i.e. candidate solution) also incorporated an associated self-adaptive mutational vector used to determine the random variation imposed during the generation of offspring networks (described below). Each of these self-adaptive parameters was initialized to a value of 0.01. Each weight set was evaluated based on how well the ANN classified the 111 available training patterns, where a diagnosis of malignancy was assigned a target value of 1.0 and a benign condition was assigned a target of 0.0. The performance of each network was determined as the sum of the squared error between the output and the target value taken over the 111 available patterns.

After evaluating all existing (parent) networks, the 250 weight sets were used to generate 250 offspring weight sets (one offspring per parent). This was accomplished in a two-step procedure. For each parent, the self-adaptive parameters were updated as:

$$\sigma'_i = \sigma_i \exp\left(\tau N(0,1) + \tau' N_i(0,1)\right) \qquad (1)$$

where $\tau = \frac{1}{\sqrt{2n}}$, $\tau' = \frac{1}{\sqrt{2\sqrt{n}}}$, $N(0,1)$ is a standard normal random variable sampled once for all 33 parameters of the vector σ, and $N_i(0,1)$ is a standard normal random variable sampled anew for each parameter. The settings for τ and τ' have been demonstrated to be fairly robust [9]. These updated self-adaptive parameters were then used to generate new weight values for the offspring according to the rule:

$$x'_i = x_i + \sigma'_i C \qquad (2)$$

where C is a standard Cauchy random variable (determined as the ratio of two independent standard Gaussian random variables). The Cauchy mutation allows for a significant chance of generating saltations but still provides a reasonable probability that offspring networks will reside in proximity to their parents. All of the offspring weight sets were evaluated in the same manner as their parents.

Selection was applied to eliminate half of the total parent and offspring weight sets based on their observed error performance. A pairwise tournament was conducted where each candidate weight set was compared against a random sample from the population. The sample size was chosen to be 10 (a greater sample size indicates more stringent selection pressure). For each of the 10 comparisons, if the weight set had an associated classification error score that was lower than the randomly sampled opponent it received a "win." After all weight sets had participated in this tournament, those that received the greatest number of wins were retained as parents of the next generation. This process affords a probabilistic selection, not unlike that achieved in annealing methods [13], allowing for the possibility of climbing up and out of hills and valleys on the error response surface.

This process was iterated for 200 generations, whereupon the best available network as measured by the training performance was used to classify the held-out input feature vector. The result of this classification was recorded (i.e., the output value of the network and the associated target value) and the process was restarted by replacing the held-out vector and removing the next vector in succession until all 111 patterns had been classified. Note that each final classification was made using a network that was not trained on the pattern in question.

Each complete cross validation was repeated 16 times, with different randomly selected populations of initial weights, to determine the reliability of the overall procedure. A typical rate of optimization in each training run is shown in Figure 1. The probability of detection, P(D), and false positive, P(FP), vary with the discrimination threshold applied to the output of the networks. As the threshold value is lowered, the network can correctly identify a greater number of cancers, but this comes at the expense of a higher false positive rate. Conversely, the false positive rate can be lowered by raising the threshold value, but this in turn decreases the sensitivity of the procedure.

Results and Discussion

The effectiveness of the classification procedures can be assessed using receiver operating characteristic (ROC) analysis, where the probability of detecting a malignancy is traded off as a function of the likelihood of a false positive. A typical ROC curve for the 16 trials is offered in Figure 2. The area under the curve, typically denoted A_Z, provides a useful measure for assessing the performance of the system. The mean area $\bar{A}_Z$ (determined using polynomial splines) was 0.8982 with a standard error of $s_{\bar{A}_Z} = 0.0098$. The best network achieved $A_Z = 0.9345$.

The average performance of the evolved ANNs in terms of A_Z is comparable to that of [14, 15], which also used mammographic features interpreted by a radiologist. The ANN in [15], which used 18 input features (both radiographic and clinical) and possessed 10 hidden nodes, yielded a specificity of 0.62 at a sensitivity of 0.95. By comparison, radiologists attained only a 0.3 specificity on the same data. The evolved networks in the current study yielded a mean specificity of 0.6187±0.0285 at 0.95 sensitivity. Although this result is almost identical to the performance offered in [15], the evolved networks are more parsimonious models (about an order of magnitude fewer degrees of freedom), and may therefore offer greater generalizability while requiring less computational effort.

One criticism of the use of ANNs in medical diagnoses is that they are black box methods, and in general are not "explainable" [16]. The success of small ANNs in diagnosing breast cancer, as observed here, offers the promise that suitable explanations for the network's behavior can be induced, perhaps leading to a greater acceptance by physicians and ultimately a useful tool.

It would appear that training by simulated evolution or other stochastic methods is key to developing these parsimonious networks. Under the more common gradient-based training method of error back propagation, the search for appropriate ANN weight sets can stagnate at local optima. These can be overcome by adding additional nodes and weights, but the resulting networks are no longer as parsimonious as may be possible. Evolutionary algorithms offer the potential for overcoming multiple optima on the error response surface, as well as simultaneously adjusting the ANN topology. Further, the evolutionary training method can be used regardless of the payoffs for correct and incorrect classifications, which may be important in trading off the costs of sensitivity and specificity.

Acknowledgments

The authors thank Peter Angeline, Wirt Atmar, Sal Cianci, Gary Fogel, and Larry Fogel for their comments on the paper. This effort is sponsored by the Department of the Army, Grant No. DAMD17-96-1-6312. The content of the information presented does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

References

[1] Boring, C.C., Squires, T.S. and Tong, T. (1993) Cancer statistics. CA: Cancer Journal for Clinicians, 43, 7-26.

[2] Strax, P. (1989) Make Sure that You do not have Breast Cancer, St. Martin's, NY.

[3] Elmore, J.G., Wells, C.K., Lee, C.H., Howard, D.H., and A.R. Feinstein (1994) Variability in radiologists' interpretations of mammograms. N. England J. Med., 331, 1493-1499.

[4] Ciccone, G., Vineis, P., Frigerio, A., and Segnan, N. (1992) Inter-observer and intra-observer variability of mammogram interpretation: a field study. Eur. J. Cancer, 28A, 1054-1058.

[5] Haykin, S. (1994) Neural Networks. Macmillan, NY.

[6] Giger, M. and MacMahon, H. (1996) Image processing and computer-aided diagnosis. Imaging and Information Management: Computer Systems for a Changing Health Care Environment, 34:3, 565-596.

[7] Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M.A., Adler, D.D., and Goodsitt, M.M. (1996) Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. IEEE Trans. Medical Imaging, 15:5, 598-610.

[8] Porto, V.W., Fogel, D.B., and Fogel, L.J. (1995) Alternative neural network training methods. IEEE Expert, 10, 16-22.

[9] Fogel, D.B. (1995) Evolutionary Computation. IEEE Press, NY.

[10] Bäck, T. (1996) Evolutionary Algorithms in Theory and Practice. Oxford, NY.

[11] Bäck, T., Fogel, D. B., and Michalewicz, Z. (eds.) Handbook of Evolutionary Computation. Oxford, NY.

[12] Floyd, C.E., Lo, J.Y., Yun, A.J., Sullivan, D.C., and Kornguth, P.J. (1994) Prediction of breast cancer malignancy using an artificial neural network. Cancer, 74, 2944-2998.

[13] Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983) Optimization by simulated annealing. Science, 220, 671-680.

[14] Wu, Y., Giger, M.L., Doi, K., Vyborny, C.J., Schmidt, R.A., and Metz, C.E. (1993) Application of neural networks in mammography: Applications in decision making in the diagnosis of breast cancer. Radiology, 187, 81-87.

[15] Baker, J.A., Kornguth, P.J., Lo, J.Y., Williford, M.E., and Floyd, C.E. (1995) Breast cancer: Prediction with artificial neural networks based on BI-RADS standardized lexicon. Radiology, 196, 817-822.

[16] Kahn, C.E. (1996) Decision aids in radiology. Imaging and Information Management: Computer Systems for a Changing Health Care Environment, 34:3, 607-628.

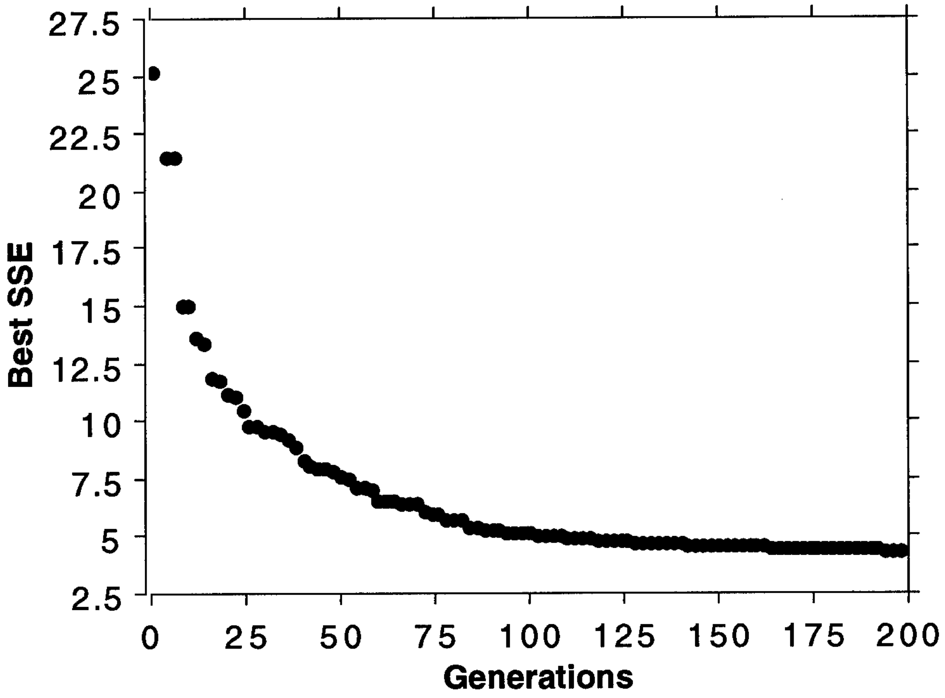
Table I. The features and rating system used for assessing mammograms in the current study. Assessment was made by the domain expert (Wasson).

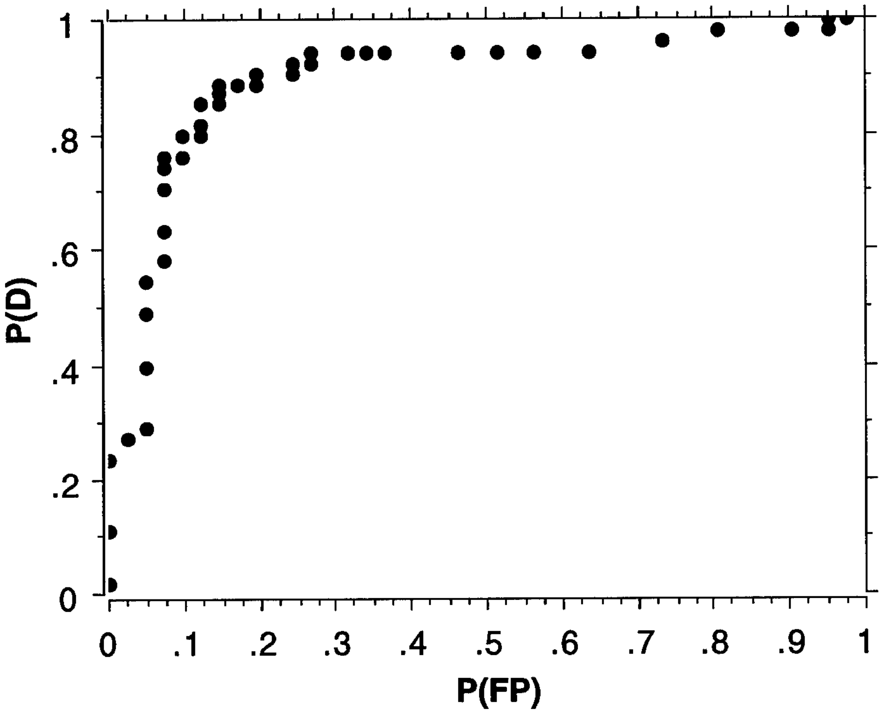
1. Mass size: either zero or in mm.
2. Mass margin: (each subparameter rated as none (0), low (1), medium (2), or high (3))
    (a) Well circumscribed
    (b) Microlobulated
    (c) Obscured
    (d) Indistinct
    (e) Spiculated
3. Architectural distortion: none or distortion
4. Calcification number: none (0), < 5 (1), 5-10 (2), or > 10 (3).
5. Calcification morphology: none (0), not suspicious (1), moderately suspicious (2), or highly suspicious (3)
6. Calcification density: none (0), dense (1), mixed (2), faint (3)
7. Calcification distribution: none (0), scattered (1), intermediate (2), clustered (3)
8. Asymmetric density: either zero or in mm.

Figure Captions

Figure 1. Typical optimization performance using simulated evolution to train the ANN. The graph depicts the sum of squared error (SSE) of the best network in the population as a function of the number of generations. Training was performed over 111 patterns, with one pattern held out for testing in cross validation. The sufficiency of the number of generations is indicated as the learning curve approaches an asymptote.

Figure 2. A typical ROC curve (raw data) generated in one complete cross validation where each of 112 patterns was classified in turn, based on training over the remaining 111 patterns. Each point represents the probability of detection, P(D), and probability of false positive, P(FP), that is attained as the threshold for classifying a result as malignant is increased systematically over [0,1].


27.5
25
22.5
20
17.5
15
12.5
10
7.5
5
2.5
Best SSE
0
25
50
75
100
125
150
175
200
Generations


1
.8
.6
.4
.2
0
P(D)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
P(FP)

Paper in review at *IEEE Trans. Med. Imaging*, 1997

Evolving Linear and Neural Models for Classifying Breast Masses

David B. Fogel
Natural Selection, Inc.
3333 N. Torrey Pines Ct., Suite 200
La Jolla, CA 92037
dfogel@natural-selection.com
(619) 455-6449 tel
(619) 455-1560 fax
(corresponding author)

Eugene C. Wasson
Maui Memorial Hospital
221 Mahalani
Wailuku, HI 96793
wasson@maui.net

Edward M. Boughton
Hawaii Industrial Laboratory, Inc.
P.O. Box 1275
Wailuku, HI 96793
boughton@maui.com

Vincent W. Porto
Natural Selection, Inc.
3333 N. Torrey Pines Ct., Suite 200
La Jolla, CA 92037
bporto@natural-selection.com

Summary

Computational methods can be used to provide a second opinion in medical settings and may improve the sensitivity and specificity of diagnoses. In the current study, evolutionary programming is used to train artificial neural networks and linear discriminant models to detect breast cancer in suspicious masses using radiographic features and patient age. Results on 139 suspicious breast masses (79 malignant, 60 benign, biopsy proven) indicate that a significant probability of detecting malignancies can be achieved at the risk of a small percentage of false positives.

Keywords: Breast cancer, computer-assisted diagnosis, artificial neural networks, evolutionary programming

Supported by the Department of the Army, Grant No. DAMD17-96-1-6312. The content of the information presented does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

Introduction

Carcinoma of the breast is second only to lung cancer as a tumor-related cause of death in women. There are now more than 180,000 new cases and 45,000 deaths annually in the United States alone [1]. Intra- and inter-observer disagreement and inconsistencies in mammographic interpretation [2, 3] have led to an interest in using computerized pattern recognition algorithms, such as artificial neural networks (ANNs) [4] and linear discriminant models (also known as linear classifiers), to assist the radiologist in the assessment of mammograms. The "second opinion" offered by a reliable automated system may be useful in reducing false-negative diagnoses [5, 6], and other oversights that may result from poor mammographic image quality, physician fatigue, or alternative sources. It may also serve to improve the accuracy of determining those patients for whom further assessment and possible biopsy is indicated.

This note reports on the use of evolutionary programming (EP) [7] to train ANNs and linear classifiers to assess 139 suspicious breast masses based on 12 radiographic features and patient age. Evolutionary optimization is suggested for training ANNs because the error between the actual output of the network and the desired target output defines a potentially multimodal response surface over a multidimensional hyperspace (the dimension is equal to the number of weights). Commonly employed methods for finding weight sets (e.g., error back propagation) are essentially gradient methods and are subject to entrapment in locally optimal solutions. The resulting weight sets are often unsuitable for practical applications. In contrast, evolutionary algorithms avoid entrapment in local optima, and there are several mathematical proofs that variations of these procedures provide asymptotic global convergence, rather than merely local convergence [8]. Moreover, there is empirical evidence that the methods are robust to many difficulties in possible response surfaces, including multiple minima or maxima, constraints, disjoint feasible regions, and random perturbations [8]. The current research extends previously reported results [9] in which EP was used to train ANNs for such classification on a smaller data base. Readers unfamiliar with the use of ANNs for breast cancer detection are referred to [5, 6, 9].

Method

Data were collected by assessing film screen mammograms using a set of 12 radiographic features as determined by the domain expert (Wasson) (Table I). The features selected paralleled those offered in [10], with some modifications to increase the orthogonality of the features, as well as the inclusion of patient age. These features were assessed in 139 cases of suspicious breast mass, all of which were examined by open surgical biopsy with the associated pathology indicating whether or not a malignant condition had been found. In all, 79 cases were associated with a biopsy-proven malignancy, while 60 cases were indicated to be negative by biopsy.

These data were processed using two input-output models: (1) a simple feedforward ANN restricted to two hidden sigmoid nodes (following the maxim of parsimony, this being the most simple architecture that can take advantage of the nonlinear properties of the nodes), with a single linear output node, resulting in 31 adjustable weights, and (2) a linear classifier (created by reducing the number of hidden nodes in the above ANN to one). Evolutionary programming was used to train both the ANNs and the linear discriminant classifier in a leave-one-out cross validation procedure[1]. Specifically, for each complete cross validation where each sample pattern was held out for testing and then replaced in a series of 139 separate training procedures, a population of 250 models of the chosen architecture (ANN or linear) was initialized at random by sampling weight values from a uniform random variable distributed over [-0.5,0.5]. Each weight set (i.e. candidate solution) also incorporated an associated self-adaptive mutational vector used to determine the random variation imposed during the generation of offspring networks (described below). Each of these self-adaptive parameters was initialized to a value of 0.01. Each weight set was evaluated based on how well the model classified the 138 available training patterns (with the remaining pattern held out for testing), where a diagnosis of malignancy was assigned a target value of 1.0 and a benign condition was assigned a target of 0.0. The performance of each network was determined as the sum of the squared error between the output and the target value taken over the 138 available patterns. Note that this weights all errors and correct classifications equally. This restriction need not hold in practice.

After evaluating all existing (parent) networks, the 250 weight sets were used to generate 250 offspring weight sets (one offspring per parent). This was accomplished in a two-step procedure. For each parent, the self-adaptive parameters were updated as:

$$\sigma'_i = \sigma_i \exp\left(\tau N(0,1) + \tau' N_i(0,1)\right) \qquad (1)$$

where $\tau = \frac{1}{\sqrt{2n}}$, $\tau' = \frac{1}{\sqrt{2\sqrt{n}}}$, $N(0,1)$ is a standard normal random variable sampled once for all 31 parameters of the vector σ, and $N_i(0,1)$ is a standard normal random variable sampled anew for each parameter. The settings for τ and τ' have been demonstrated to be fairly robust [8]. These updated self-adaptive parameters were then used to generate new weight values for the offspring according to the rule:

$$x'_i = x_i + \sigma'_i C \qquad (2)$$

where C is a standard Cauchy random variable (determined as the ratio of two independent standard Gaussian random variables). The Cauchy mutation allows for a significant chance of generating saltations but still provides a reasonable probability that offspring networks will reside in proximity to their parents. All of the offspring weight sets were evaluated in the same manner as their parents.

Selection was applied to eliminate half of the total parent and offspring weight sets based on their observed error performance each generation. A pairwise tournament was conducted where each candidate weight set was compared against a random sample from the population. The sample size was chosen to be 10 (a greater sample size indicates more stringent selection pressure). For each of the 10 comparisons, if the weight set had an associated classification error score that was lower than the randomly sampled opponent it received a "win." After all weight sets had participated in this tournament, those that received the greatest number of wins were retained as parents of the next generation. This process affords a probabilistic selection, not unlike that achieved in annealing methods [11], allowing for the possibility of climbing up and out of hills and valleys on the error response surface.

This process was iterated for 200 generations, whereupon the best available network as measured by the training performance was used to classify the held-out input feature vector. The result of this classification was recorded (i.e., the output value of the network and the associated target value) and the process was restarted by replacing the held-out vector and removing the next vector in succession until all 138 patterns had been classified. Note that each final classification was made using a model that was not trained on the pattern in question.

Each complete cross validation was repeated 16 times for the ANNs and 10 times for the linear classifier, with different randomly selected populations of initial weights, to determine the reliability of the overall procedure. Typical rates of optimization in each training run are shown in Figure 1. The probability of detection, P(D), and false positive, P(FP), vary with the discrimination threshold applied to the output of the models. As the threshold value is lowered, the models can correctly identify a greater number of cancers, but this comes at the expense of a higher false positive rate. Conversely, the false positive rate can be lowered by raising the threshold value, but this in turn decreases the sensitivity of the procedure.

Results

The effectiveness of the classification procedures can be assessed using receiver operating characteristic (ROC) analysis, where the probability of detecting a malignancy is traded off as a function of the likelihood of a false positive. Typical ROC curves for the 16 trials involving ANNs and 10 trials with linear classifiers is offered in Figure 2. The area under an ROC curve, often denoted A_Z, provides a useful measure for assessing performance. The mean area $\bar{A}_Z$ (determined using polynomial splines of maximum 9th order) for the ANNs was 0.9290 with a standard error of $s_{\bar{A}_Z} = 0.0052$. The best network achieved $A_Z = 0.9657$. The mean area $\bar{A}_Z$ for the linear classifiers was 0.9187 with a standard error of $s_{\bar{A}_Z} = 0.0048$. The best linear classifier achieved $A_Z = 0.9397$. A *t*-test indicated no statistically significant difference ($P > 0.1$) in performance between the ANNs and linear classifiers (note that the assumptions for the test may not hold given the sample sizes)

The average performance of the evolved ANNs and linear classifiers in terms of A_Z is slightly better than observed in [9] and comparable to that of [11, 12], which also used mammographic features interpreted by a radiologist. The ANN in [12], which used 18 input features (both radiographic and clinical) and possessed 10 hidden nodes, yielded a specificity of 0.62 at a sensitivity of 0.95. By comparison, radiologists attained only a 0.3 specificity on the same data. The evolved ANNs and linear classifiers in the current study yielded mean specificities of 0.7289±0.0346 and 0.6610±0.0413, respectively. Again, these values do not indicate a statistically significant difference ($P > 0.1$), however the performance of the evolved ANNs is statistically significantly better than that offered in [12] ($P < 0.05$). Caution must be applied in such comparisons because of the unknown differences underlying the populations from which mammograms were sampled.

Nevertheless, even if the mean specificity of the evolved classifiers was judged as being almost identical to the performance offered in [12], the evolved models are more parsimonious (about an order of magnitude fewer degrees of freedom), and may therefore offer greater generalizability while requiring less computational effort. Future effort will involve using the evolved models to classify new data not used in cross validation and assess whether or not there is any observed loss in performance.

Acknowledgments

This effort is sponsored by the Department of the Army, Grant No. DAMD17-96-1-6312. The content of the information presented does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

Footnote

1. The optimum parameters for the linear models could have been determined by direct calculation, however, using the existing evolutionary optimization software facilitated the cross validation of these models.

References

[1] C.C. Boring, T.S. Squires, and T. Tong, "Cancer statistics," *CA: Cancer Journal for Clinicians*, vol. 43, pp. 7-26, 1993.

[2] J.G. Elmore, C.K. Wells, C.H. Lee, D.H. Howard, and A.R. Feinstein, "Variability in radiologists' interpretations of mammograms," *N. England J. Med.*, vol. 331, pp. 1493-1499, 1994.

[3] G. Ciccone, P. Vineis, A. Frigerio, and N. Segnan, "Inter-observer and intra-observer variability of mammogram interpretation: a field study," *Eur. J. Cancer*, vol. 28A, pp. 1054-1058, 1992.

[4] S. Haykin, *Neural Networks,* NY:Macmillan, 1994.

[5] M. Giger and H. MacMahon, "Image processing and computer-aided diagnosis," *Imaging and Information Management: Computer Systems for a Changing Health Care Environment*, vol. 34, no. 3, pp. 565-596, 1996.

[6] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M.A. Helvie, D.D. Adler, and M.M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," *IEEE Trans. Medical Imaging*, vol. 15, no. 5, pp. 598-610, 1996.

[7] V.W. Porto, D.B. Fogel and L.J. Fogel, "Alternative neural network training methods," *IEEE Expert*, vol. 10, pp. 16-22, 1995.

[8] D.B. Fogel, *Evolutionary Computation*, NY:IEEE Press, 1995.

[9] D.B. Fogel, E.C. Wasson, E.M. Boughton, and V.W. Porto, "A step toward computer-assisted mammography using evolutionary programming and neural networks," *Cancer Letters*, in press, 1997.

[10] C.E. Floyd, J.Y. Lo, A.J. Yun, D.C. Sullivan, and P.J. Kornguth, "Prediction of breast cancer malignancy using an artificial neural network," *Cancer*, vol. 74, pp. 2944-2998, 1994.

[11] Y. Wu, M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schmidt, and C.E. Metz, "Application of neural networks in mammography: Applications in decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81-87, 1993.

[12] J.A. Baker, P.J. Kornguth,J.Y. Lo, M.E. Williford, and C.E. Floyd, "Breast cancer: Prediction with artificial neural networks based on BI-RADS standardized lexicon," *Radiology*, vol. 196, pp. 817-822, 1995.

Table I. The features and rating system used for assessing mammograms in the current study. Assessment was made by the domain expert (Wasson).

1. Mass size: either zero or in mm.
2. Mass margin: (each subparameter rated as none (0), low (1), medium (2), or high (3))
    (a) Well circumscribed
    (b) Microlobulated
    (c) Obscured
    (d) Indistinct
    (e) Spiculated
3. Architectural distortion: none or distortion
4. Calcification number: none (0), < 5 (1), 5-10 (2), or > 10 (3).
5. Calcification morphology: none (0), not suspicious (1), moderately suspicious (2), or highly suspicious (3)
6. Calcification density: none (0), dense (1), mixed (2), faint (3)
7. Calcification distribution: none (0), scattered (1), intermediate (2), clustered (3)
8. Asymmetric density: either zero or in mm.

Figure Captions

Figure 1. Typical optimization performance using simulated evolution to train (a) the ANN, (b) the linear classifier (an ANN with only one hidden node). The graphs depict the mean squared error (MSE) per pattern for the best model in the population as a function of the number of generations. Training was performed over 138 patterns, with one additional pattern held out for testing in cross validation. The sufficiency of the number of generations is indicated as the learning curves approach an asymptote.

Figure 2. Typical ROC curves (raw data) generated for (a) the ANN and (b) the linear classifier in one complete cross validation where each of 139 patterns was classified in turn, based on training over the remaining 138 patterns. Each point represents the probability of detection, P(D), and probability of false positive, P(FP), that is attained as the threshold for classifying a result as malignant is increased systematically over [0,1].

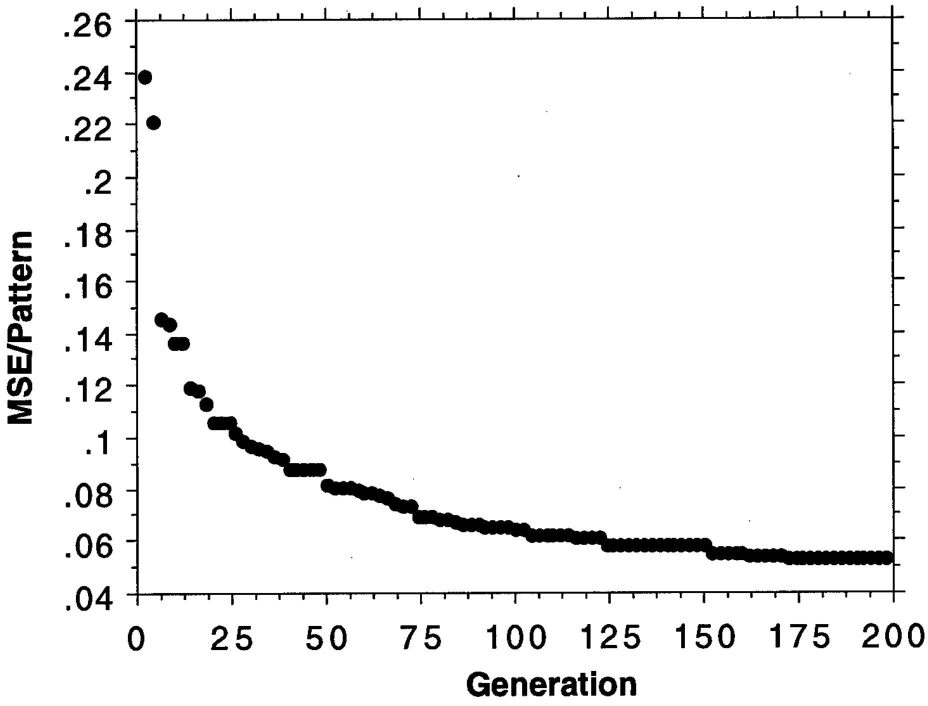

MSE/Pattern
.26
.24
.22
.2
.18
.16
.14
.12
.1
.08
.06
.04
0
25
50
75
100
125
150
175
200
Generation

Fig. 1a

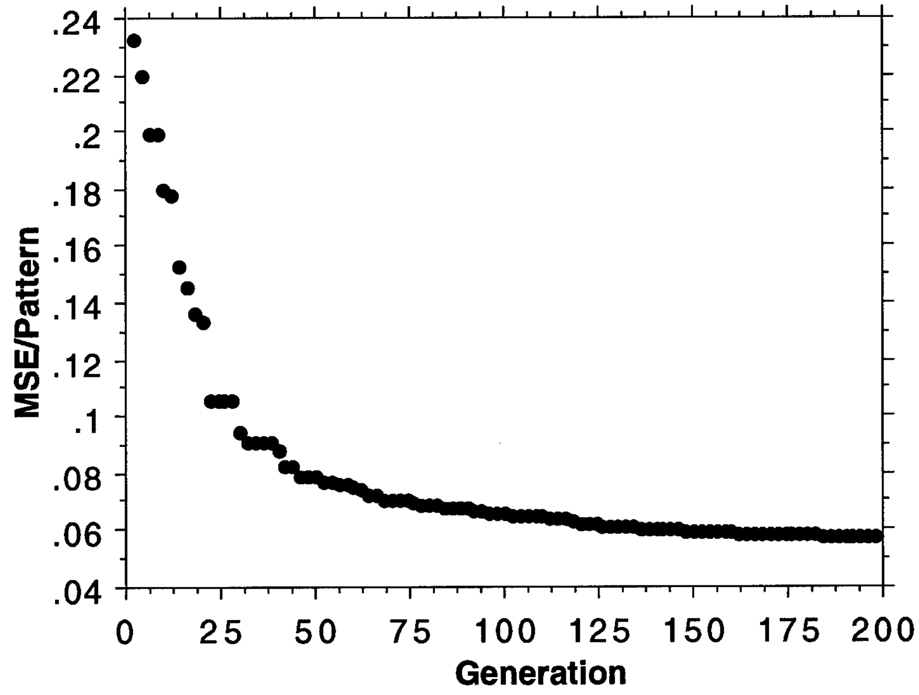

MSE/Pattern
.24
.22
.2
.18
.16
.14
.12
.1
.08
.06
.04
0
25
50
75
100
125
150
175
200
Generation

Fig. 1b

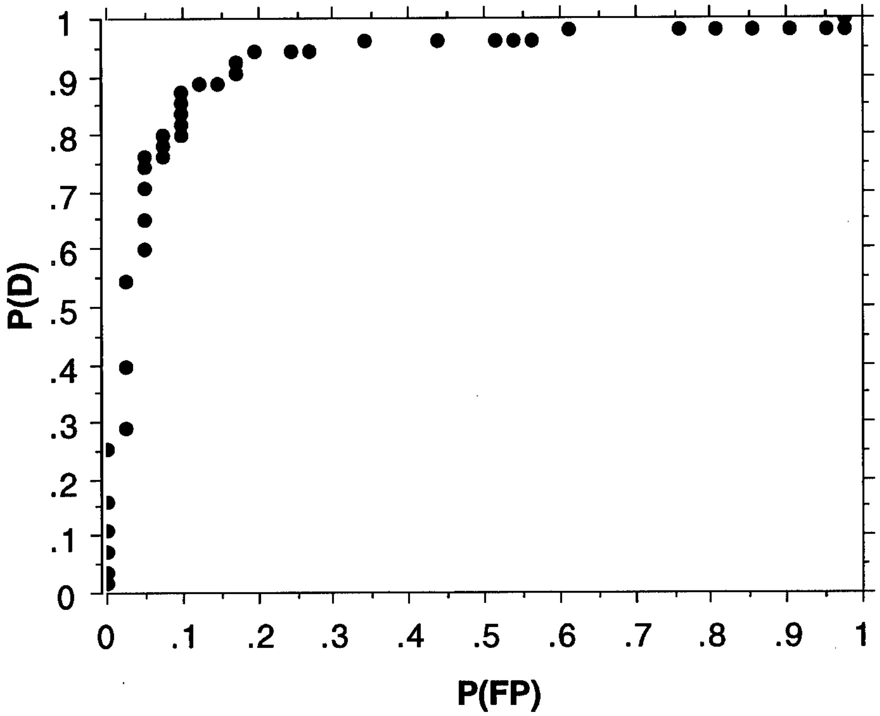


Figure 2a

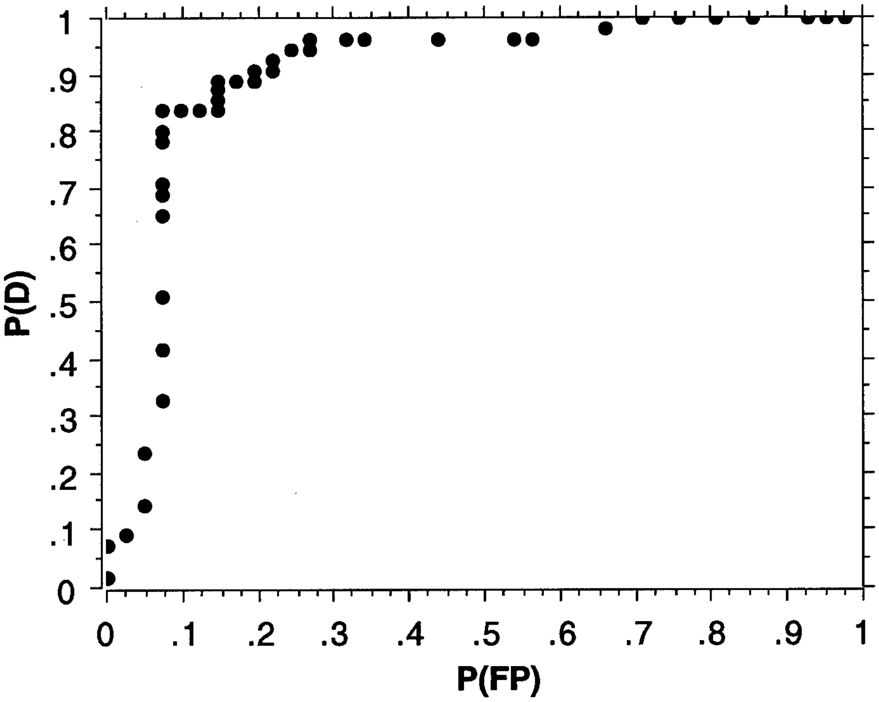


Figure 2b